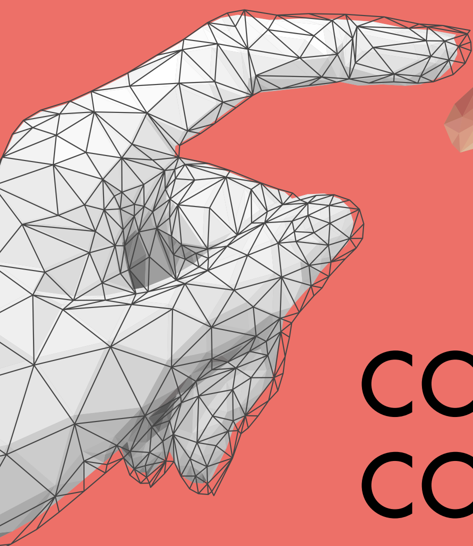


COMPENDIUM COMPUTATIONAL THEOLOGY

Vol. 1: Introducing Digital Humanities to Theology

Edited by
Christopher A. Nunn &
Frederike van Oorschot

UNIVERSITÄTS-
BIBLIOTHEK
HEIDELBERG

Compendium of Computational Theology
Vol. 1: Introducing Digital Humanities to Theology

Christopher A. Nunn and Frederike van Oorschot (Eds.)

Compendium of Computational Theology

Vol. 1: Introducing Digital Humanities to Theology

UNIVERSITÄTS-
BIBLIOTHEK
HEIDELBERG

The *Compendium Computational Theology* was created in the context of TheoLab, a research network for Computational Theology based at Heidelberg University.
All chapters have been peer-reviewed *(Single-Blind Pre-Publication Peer Review).*

The compendium was produced with the kind support of the following institutions:

- DHd – Association for Digital Humanities in the German Speaking Areas
- FEST Institute for interdisciplinary research
- Heidelberg Center for Digital Humanities (HCDH) as part of the Excellence Strategy of the German Federal and State Governments

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at https://dnb.dnb.de.

This book is published under the Creative Commons Attribution License CC BY-SA 4.0. The cover is subject to the Creative Commons License CC BY-ND 4.0.

UNIVERSITÄTS-
BIBLIOTHEK
HEIDELBERG

Published by heiBOOKS, 2024

Heidelberg University/Heidelberg University Library
heiBOOKS
Grabengasse 1, 69117 Heidelberg, Germany
https://books.ub.uni-heidelberg.de/heibooks
e-mail: ub@ub.uni-heidelberg.de

The electronic open access version of this work is permanently available on the e-book platform of Heidelberg University Library heiBOOKS: https://books.ub.uni-heidelberg.de/heibooks
urn: urn:nbn:de:bsz:16-heibooks-book-1521-1
doi: https://doi.org/10.11588/heibooks.1521

Text © 2024, Christopher A. Nunn, Frederike van Oorschot

Cover illustration: Adobe Stock, lidiia
Typesetting: text plus form, Dresden

ISBN 978-3-911056-19-9 (PDF)

Table of Contents

V. DISSEMINATION

Foreword

Since 2019, the TheoLab has concentrated on research activities at the intersection of Theology and the Digital Humanities. As the only research network working at this intersection, the conferences, workshop, reports and colloquia have developed into an important forum for presenting current research projects in both fields. Increasingly, the TheoLab has also become a central location for exciting new projects and digital initiatives within Theology. Even more, Christopher Nunn and Frederike van Oorschot have shaped the research field of "Computational Theology" and contributed heavily to the international debates on Digital Theology.

This Compendium meets two separate criteria for the intersection of Theology and the Digital Humanities. The Compendium first provides – especially the first volume – an introduction to current research initiatives within the Digital Humanities, which could be of interest for all scholars interested in the field, not only those working within Theology. For another, the second volume offers the first exploration of current computer-assisted projects in the theological disciplines and presents possible further applications. These volumes thus not only further distinguish the Computational Theology developed in Heidelberg; they also show the correlated interdisciplinary interfaces and further research perspectives for these two exciting fields of scholarship.

We are delighted that the work of the TheoLab has reached a further milestone with the publication of this Compendium and hope it initiates productive discussion and research.

Prof. Dr. Michael Gertz (Computer Science, Heidelberg)
Prof. Dr. Friederike Nüssel (Systematic Theology/Ecumenical Studies, Heidelberg)
Prof. Dr. Helmut Schwier (New Testament and Practical Theology, Heidelberg)

TheoLab Advisory Board

Foreword, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 9. DOI: https://doi.org/10.11588/heibooks.1521.c21934

I. INTRODUCTION

Compendium of Computational Theology – Introduction

Christopher A. Nunn[a] and Frederike van Oorschot[b]

[a] https://orcid.org/0000-0001-7208-8636, [b] https://orcid.org/0000-0003-4359-8949

Abstract This introductory essay provides an overview of the topics covered in the Compendium. Despite a long-standing tradition, research at the interface of theology and Digital Humanities has yet to gain traction. One reason for this research gap is the lack of infrastructure and foundational texts. Another reason is the various understandings of the task of Digital Theology. Thus, this article bases the field of Computational Theology firmly within the canon of the Digital Humanities. The second part of this essay provides insight into the genesis of this compendium and outlines the individual essay contributions in the first volume.*

Keywords Computational Theology, Digital Humanities, Interdisciplinarity, Philosophy of Science

1. Computational Theology?

The aim of this compendium is to provide an orientation to the fields of theology and Digital Humanities (DH).[1] The volume participates in two debates, the form and concept of which remain vast and unclear. With respect to theology, the compendium focuses on the small, tangible areas of academic theology in the diversity of the theological disciplines. Clarifying what is meant by the concept of DH is more difficult. This undertaking is anything but trivial, given several definitions have surfaced even within their limited history (on the origin of DH, see Piotrowski in this volume, pp. 33–35). Kirschenbaum (2010,1) writes: "'What is Digital Humanities?' essays like this one are already genre pieces."[2] As a working definition of DH, Sahle's (2017, 9) concept will suffice:

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

1 In order to facilitate the widest possible dissemination of this orientation, all chapters are presented in both German and English. In the translations, quotations have also been translated into the target language by the editorial team.

2 Cf. https://whatisdigitalhumanities.com (accessed on 18 May 2024), where Jason Heppler lists 817 attempts to define DH

Christopher A. Nunn and Frederike van Oorschot: Compendium of Computational Theology – Introduction, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 13–28. DOI: https://doi.org/10.11588/heibooks.1521.c21936

> Essentially, however, DH is about the development, use, and critical reflection of digital processes in the humanities. DH take up the issues of the humanities and combine them with solutions from computer science and, in some cases, other disciplines – for example, advanced imaging methods from engineering, geoinformation systems from geography, empirical methods from the social sciences, or information theory approaches from *library and information science.*

According to Rapp (2021, 8), an increase in the activities of DH research can be observed, which have been politically driven by specific funding programs at prominent funding institutions. The social expectations associated with this increase are illustrated by Schmale's (2016, 299) research on musicology:

> Musicology is no different than other disciplines in the humanities and cultural sciences: any discipline without at least one digital branch is currently under pressure to justify itself. In the digital humanities, musicology is catching up with other disciplines: degree courses with a focus on digital musicology, PhD programs, professorships, digital projects, conferences – pulling out all stops.

Can the "development, use, and critical reflection of digital processes" also be seen as increasing in the research practices of academic theology and its disciplines? According to Hutchings & Clivaz (2021, 6), there is without a doubt a "digital branch" (if not even an entire forest landscape) in theology:

> Christian theology, religious studies and biblical studies have a long, rich and productive history of interaction with the academic digital humanities. There is no unique Christian way to do DH, but the numerous signs of academic institutionalization of DH, the rich libraries of academic DH publications and the extraordinary global Christian interest in digital theology and digital Bible study all demonstrate the value of an introductory book to Christianity and the digital humanities (Hutchings & Clivaz 2021, 6).

Other voices are a bit more restrained. Theologians Heyden & Schröder, for example, refer to "pioneering achievements" in theology in the DH and thus confirm a "long, rich and productive history of interaction with the academic digital humanities."[3]

3 Like many others (e.g., Peters in this volume on p. 316), they refer to the preliminary work of *Index Thomisticus*, which began with conversations in the 1940s between the Jesuit priest Roberto Busa and the IBM founder Thomas Watson Sr. On the origins of the *Index Thomisticus* as the "founding myth of DH," see Thaller (2017, 3). For a monographic review of its origins, see Jones (2016). Despite the merits of Busa, however, it must be noted that the timing (cf. Blaney [2021, 7]) and

Fig. 1 Spheres of the Digital Humanities

At the same time, they observe that "aside from the pioneering achievements [...], the breadth of theology [...] has reacted hesitantly to the 'digital turn.'" This observation is reflected in studies observing the academic domains of DH. Theology was not included among the 15 "disciplines that are related to DH in some way," that Luhmann & Burghardt (2021, 150) consider in their scientometric study on DH and the academic

working practices (cf. Nyhan [2023, 23]) of the project are not uncontroversial. The disciplinary classification can also vary. Krämer (2019, 244), for example, explains Busa's original concern in his doctoral thesis as "a genuinely philosophical one" and concludes: "It was precisely a philosophical [!] research project that became the pioneer of the Digital Humanities at the beginning of the 1950s."

disciplines. Nor is theology mentioned in the oft cited spherical model (fig. 1) by Sahle (2013, 6).

Even from these limited observations, one can confirm: theologians are largely absent from the DH-Community.[4] However, the current lack of involvement does not indicate that theology cannot use digitally supported research. When confronted with the question of digitization of research material, church historian Hubert Wolf elaborates "that with the currently available digital editions, students can use certain keyword searches in their studies to work on term papers that would otherwise require a large amount of time in the archive or on which a doctoral student would have worked for long time" (Burke & Hiepel 2021, 20). In fact, there are large databases and digital collections in theology (on this exemplary selection, see A. von Stockhausen in this volume, pp. 94–95). This collection of resources is summarized in a working paper from *Digital Research Infrastructure for the Arts and Humanities* (DARIAH): "Rapid progress is possible in theology because a large number of primary texts already exist in digital form" (Reiche et al. 2014, 21).

The digitization of resources is not the end of digital research, but the beginning, particularly within the context of DH: the potential of DH is not exhausted in the retroactive digitization of theological works (for a critique of this position, see Zahnd 2020, 117). Theological projects that go beyond this simple approach, pursuing questions unanswerable without the use of information technological resources, have become increasingly prevalent (see the chapter by Nunn in this volume); however, these projects are limited when compared to other humanities disciplines.

Through the events of the TheoLab, a research network founded in 2019 and led by the authors of this introduction,[5] which is dedicated to precisely such questions in colloquia for young researchers, workshop reports and conferences, it quickly became clear that there are either major reservations among theologians about getting involved in DH, or at least uncertainties regarding meaningful research questions and possibilities for technical implementation. This hesitation is partly due to the lack of specific infrastructures and foundational works, and partly to theoretical, methodological, and epistemological reservations (van Oorschot 2021).

There is a clear push towards institutionalization in other humanities disciplines, such as in the field of history, where, since 2009, several professorships, conference series, journals, and courses have been created using the term *Digital History* (for the first use by Peter Haber, see Döring et al. [2022, 5]). Salmu (2021, 7) describes this paradigm:

4 Nunn (2024) presents a number of potential explanations for the apparent absence of theologians within the DH-Community.

5 It would be remiss not to mention Stefan Karcher, who founded the TheoLab in 2019 and played a leading role in its organisation until his departure from Heidelberg in 2021. Similarly, Selina Fucker has been involved in TheoLab for a number of years.

> Digital history acknowledges its origins in historians' serious efforts to engage with the Internet, digital tools and information technology. It is also a repository for the computational methods that were developed in the digital humanities and can be applied in and refined for solving historical problems. The definition of digital history can today be reformulated as follows: digital history is an approach to examining and representing the past; it uses new communication technologies and media applications and experiments with computational methods for the analysis, production and dissemination of historical knowledge.

There are similar definitions of Digital Classics (cf. Schubert 2015, 1), Digital Philology (cf. Adler et al. 2020, 1), Digital Philosophy (cf. Gramelsberger 2023, 111) or Digital Art History (cf. Schelbert 2018, 42).

Given the range of meanings of theology, simply adding Digital Theology to the list of disciplines would not suffice. According to van Oorschot (2023, 25), there are four areas that can be addressed under the label of Digital Theology:

1. Theology with digital resources, tools, or methods. This includes approaches to doing theology with digital methods, such as in the adaptation of the Digital Humanities.
2. Theology in digital spaces. This describes attempts to do theology using digital media, such as theological podcasts, blogs, or online journals.
3. Theological reflection on digitalization. Reflection on digital change and correlated topics can currently be found in the fields of practical theology (e. g. educational theory) or ethics (e. g. use of artificial intelligence, cyberwar).
4. The digital transformation of theology. Another approach is the question of how categories, thought models, and questions of theological reflection themselves are changed by the process of digitalization.

The first mentioned dimension, which aligns with the understanding of other humanities disciplines has rarely been the research focus. In theology, attention and the associated resources have been devoted to other areas. *The Global Network for Digital Theology* (GoNeDigiTal), which Hutchings & Clivaz primarily have in mind in the citation above, describes itself as "an international network that connects those involved in research and reflection on the interaction and expression of theologies within digital cultures, media, and technology."[6] The practices of the DH-Community are only marginally addressed; rather, the focus is on the media studies movement,

6 Cf. https://www.gonedigital.media/what-we-do (accessed on 19 May 2024).

to which Florian Höhne belongs as having the first professorship for Digital Theology in the German-speaking realm.[7]

The TheoLab created the first ever infrastructure entirely dedicated to theology as DH discipline.[8] The *Compendium Computational Theology* provides the first comprehensive work published that is specifically tailored to the needs of theological researchers in the DH.[9] Computational Theology thus indicates "theological questions that are investigated with the help of computational approaches" (van Oorschot 2023, 29). The term is borrowed from the Computational Humanities, whose representatives set themselves apart from the *umbrella* of DH by returning to this older term to bring the technical aspects back to the fore (see the chapter by Piotrowski in this volume; for a description of the initial situation in DH, see also Berry & Fagerjord 2017, 36 f.). With this compendium, we hope to intensify the optimistic attitude that has recently been present at the interface of theology and the DH, to contribute to ensuring that theological research remains capable of discourse from a holistic scientific perspective by providing conceptual guidance, and to develop constructively in both fields in the networking of theology and the DH.

2. The Conception of the Compendium

Theologians should not have a hard time engaging with the DH, particularly given both are very heterogenous disciplines.[10] Dalferth (2006, 5) claims, accordingly: "Protestant Theology is in no way a unified discipline; rather, theology is an ensemble of interrelated subjects and disciplines grouped around a shared task." Nüssel (2006, 92) explains this notion further:

7 On the profile of Florian Höhne, see https://www.theologie.fau.de/person/prof-dr-florian-hoehne (accessed on 19 May 2024).

8 On the task and essentials of the Digital Humanities, see https://dig-hum.de/digitale-geisteswissenschaften (accessed on 24 May 2024): "The humanities encompass a large group of disciplines that research all aspects of human society, culture, language, and history as well as thought and communication. The Digital Humanities share these research areas and endeavor to develop the processes of acquiring and communicating knowledge under the conditions of a digital and media world. To this end, the DH research and teach in areas such as the digitalization of knowledge and cultural heritage, the application and further development of digital tools, the operationalization and answering of research questions, and reflection on the methodological and theoretical foundations of the humanities in a digital world."

9 At the same time, this volume also offers scholars in other humanities disciplines an application-oriented introduction to various methods and practices of the DH-Community.

10 The question of whether DH is a separate discipline, an auxiliary science, or a set of methods is deliberately excluded in this context. Hamidović (2016, 2–6) offers a helpful introduction to this complex of topics.

> It was only in the 20th century that the theological disciplines developed into independent areas of research, which are only carried out by specially qualified experts and are therefore considered to be different areas of responsibility. Although the historical, systematic, and practical perspectives of theological understanding have largely diverged in research practice, and consequently also in teaching, they nevertheless belong together in substance.

Thaller (2017, 13) also emphasizes the "breadth of the research field," with regard to DH, even if there are "similarities between the different types." He continues (14): "Which information technology tools are particularly appropriate for which humanities subjects heavily depends on the self-understanding of the subject in question." The compendium hopes to provide this tool for understanding. It does not espouse either *the* theological nor *the* DH method. Depending on the theological discipline, certain approaches to DH are preferable to others. Choosing an approach that is incompatible with one's own research question costs valuable resources (cf. Krautter et al. 2023, 16). The compendium offers initial guidance in this regard.

The compendium will be divided into two volumes. The contributions in the first volume present the research practices of the DH from the perspective of the DH-community (see below). The theological authors of the second volume will be confronted with the task of exploring the potentials and limits of these practices for their respective disciplines. To this end, we organized a workshop in September 2023 in which the contributors to volume 2 were able to view and discuss an early version of the first volume.

The compendium can be used in different ways:

Scenario 1: A theology student about to graduate is considering writing a church history dissertation on early Christian martyr acts with the help of a historical network analysis. But is this approach effective? After reading the relevant chapter in volume 1, the student will be able to answer this question, as she is now aware of the potentials, pitfalls and best practices of this method.

Scenario 2: The same student asks whether the dissertation could be reasonably located in the field of Digital Humanities. After reading the chapter on ancient church history in volume 2, the student has an idea of whether it would fit and, if so, which DH practices could take her further in concrete terms. She can then take a closer look at these in the first volume.

Given the breadth of the research field in the DH and the numerous (sub)disciplines of theology, two volumes, despite their already considerable scope, will inevitably only contain a certain selection of the possible approaches at the interface of DH and theology. Naming a few examples here will suffice. E.g., the examination of three-dimensional artefacts, which represents a pivotal area of interest within the field of Christian archaeology and constitutes a core subject of inquiry within the domain of DH, is conspicuous by its absence. The use of AI could be given even greater focus in the analyses. Discourse analysis could be fleshed out with further individual methods, such as chapters on *Web Scraping*[11] or *Argument Mining.*[12] In the section on dissemination, digital forms of publication could be given their own chapter.

These two volumes should therefore only be seen as a prelude. As a *living handbook,* they will be continued online in the future, so that additions and updates to the material will still be possible after publication. In this manner, we present a comprehensive and reliable foundation that will retain its value and applicability in the years to come.

3. Overview of the Volume

3.1 Introduction

Before the practices of Computational Theology can be outlined, the specifics of the discipline must first be detailed. There are three contributions in the volume to fulfil this task. **Michael Piotrowski** introduces the relationship between Computational and Digital Humanities, showing the two disciplines are not too far apart, but are rather manifestations of two different cultural traditions. That these traditions can also be mapped onto the landscape of theological research is illustrated by **Erin Raffety,** who establishes guidelines for a Computational Theology from the perspective of a predominantly Anglo-Saxon *Digital Religion* paradigm using the example of the theological use of video games. Finally, **Christopher A. Nunn** ties the various threads together and presents interdisciplinary projects that correspond to a Computational Theology.

11 This method is used, for example, to investigate mourning practices on the Internet at the URPP Digital Religion(s) at the University of Zurich. Cf. the project website: https://www.digitalreligions.uzh.ch/de/research/internaldynamics/p1_public_valediction.html (accessed on 20 May 2024).

12 This method is currently being tested for mapping theological discourse formations using the example of “suicide” in the Department of Catholic Theology at the University of Passau. Cf. the project website: https://www.ktf.uni-passau.de/digital-methods (accessed on 20 May 2024).

3.2 Media

Christianity is a textually centered religion, which makes theological research largely centered around intensive textual work.[13] DH have also placed a significant emphasis on textual analysis. However, Kohle (2018, 16) has highlighted a potential limitation of this approach, suggesting that:

> [The Digital Humanities are] largely focused on linguistic artifacts. Generally, this focus likely has to do with the dominance of language in logocentric European culture, but more specifically with the fact that philologies are institutionally dominant in the European academic system. Images and sounds are relegated to a edges; as art media, they are dealt with in art history and musicology. Even if art history, for example, formulates a certain claim to universality with its extension from art to images, it remains institutionally marginal, or even non-existent. At this point, the situation in the digital humanities only reflects the general situation (cf. also Manovich 2020, 7).

Recently, however, the DH-community has become increasingly aware of other media. This shift can be seen, for example, by the motto of the 6th annual conference of the *Association for Digital Humanities in the German Speaking Areas* (DHd), which met in 2019 in Mainz and Frankfurt: "Digital Humanities: multimedial and multimodal."[14]

Daniel Stökl Ben Ezra discusses transcription techniques, layout analysis, and computer paleography in an essay on text digitization, thus presenting state-of-the-art of automatic analysis of manuscripts and ambiguous texts. **Hubertus Kohle's** essay then changes gears and introduces the properties of the digital image, image databases, and AI-controlled image generators on image digitization. **Christof Weiß** enters the field of computational audio and music analysis. In his essay, Weiß illustrates the potential of audio recordings for the study of church music. **Manuel Burghardt, John Bateman, Eric Müller-Budack** and **Ralph Ewerth** present an overview of computational tools and methods for film and video analysis and use the TV series "Game of Thrones" as an example to show how narrative patterns can be researched with the aid of computers.

13 A prominent example of this view of theology can be found in the DH: Moretti (2000, 57) defines his idea of *close reading* as a "theological exercise – very solemn treatment of very few texts taken very seriously." Dalferth (2018, 435), can also be mentioned here as a theological reference, according to which theology is a "text-related discipline of reflection."

14 Cf. the conference website: https://dhd2019.dig-hum.de (accessed on 21 May 2024).

3.3 Forms of Digital Text Analysis

Despite the prevalence of multimedia, the text is still the central object of theological research. We thus decided to limit the detailed analysis section to this medium. The largest possible (albeit not exhaustive!) selection of digital text analyses should help to find answers to a variety of theological research questions.

In the essay "*Python* or *R*? Getting started with programming for humanists," **William Mattingly** addresses the question of why acquiring programming skills and delving deeper into digital analysis methods is a worthy pursuit. What at first glance seems time-consuming (such as learning a programming language) can actually save research time in the end, depending on the research question. Acquiring the specific knowledge should be done in a targeted manner. Depending on the subject matter and research interest, different levels of knowledge may be required, including program packages that require no programming knowledge at all.[15]

In an essay on stylistic analysis, **Fotis Jannidis** offers the first specific approach to digital text analysis. The focus is on stylometric methods, i. e., a corpus-based analysis of style using quantitative methods that aim to assign a text to a group of other texts based on stylistic characteristics. This type of approach is often (but not exclusively) chosen to investigate questions about the authorship of a text *(authorship attribution).* Unsurprisingly, this approach also has a long tradition in theology, as Jannidis illustrates with the question of the authenticity of the Apostle Paul's letters.

Network analysis is the most well-known DH method, which creates visualizations of various relational networks (not only between specific actors or institutions, but also at the lexical level). This method has already been successfully applied in several theological disciplines. **Caitlin Burge** describes useful applications of this method based on exemplary research studies, while also mentioning potential pitfalls, such as the consequences of reductionist network visualizations, which are a result of a lack of prior consideration when conceptualizing the network.

Another important methodological approach, e. g. in practical theology, is digital discourse analysis, which is addressed by **Alexander Lasch,** who provides readers of the compendium with the necessary theoretical foundation and focuses primarily on the approaches of German discourse linguistics. Discourse analysis is about iden-

15 Cf. Dombrowski (2023, 143): "There are some kinds of DH work where coding matters. Pre-built tools will always have limitations; in their creation, developers must take decisions that constrain the kinds of questions the tool can be used to answer […]. Coding is a skill that takes time, practice, and ongoing effort to learn, but investing the energy to improve one's coding skills will not, by itself, prepare a scholar to do skillful work that is a meaningful contribution to scholarship. For that, it is more important to develop skills around the selection and preparation of data, around matching humanities questions with appropriate quantitative methods (if any exist), around carefully reading others' documentation and either producing one's own code or successfully communicating to a programmer what needs to be created – both for oneself and for future scholars to use."

tifying certain semantic relationships between different texts. Various methods can be used for this purpose. Some of these are discussed in the following chapters. The Named Entity Recognition method, which automatically recognizes entities in texts and essential for discourse analysis, is discussed in a chapter by **Evelyn Gius.** In this process, certain entities such as persons, places, data, or concepts are determined in machine processes using textual features. Since the frequency and distribution of these entities can also be read out, this method, which has not yet been used in theology, is suitable for tracing certain developments. However, discourses can also be traced using so-called topic modeling, which is addressed in the volume by **Melanie Althage.** Using various algorithms, large text corpora can be searched for specific groups of words that occur together in the individual documents with a striking statistical frequency. In this way, an overview of the content structure can be gained, which is particularly useful for making an initial hypothesis. However, it is important to be able to understand how the word groups come about and to configure the settings sensibly so as not to rely blindly on a *black box.*[16] Althage's essay can serve as an important guide for these initial findings.

However, discourses can also be visualized using more specific methods. One method is sentiment analysis, which **Rachele Sprugnoli** describes in detail. The Natural Language processing (NLP) method aims to identify and categorize opinions, emotions, and personal assessments that have been written about various entities, events, and topics. The method can, for example, help to identify the attitude of individuals to certain discourses. The main difficulties with this approach lie in recognizing irony or implicit expressions of opinion that presuppose a certain knowledge of the world. If one considers Dalferth's interpretation of theology as a text-related science of reflection, it makes sense to dedicate a chapter to digital intertextuality research. **Julia Nantke** traces the history of this strand of research and describes the methods used to evaluate intertextual relationships digitally (manually or automatically). The development of discourses, for example, could be traced on this basis. Another way of approaching texts using computer-aided methods concerns the spatial dimension. In a chapter on spatial analysis or more precisely on new literary geography, **Matthew Wilkens** describes how quantitative analyses can support qualitative studies (and thus follows a *mixed-methods* approach). He cites several impressive projects in which certain metadata from literary texts is collected and mapped in order to reveal developments in an author, an epoch, a region, or a discourse.

What happens once the respective analysis has been completed? Regarding information visualization, **Janelle Peters** guides the reader through various forms

16 Cf. van Zundert et al. (2020, 124): "The lack of knowledge about what is actually taking place in these software 'black boxes' and about how they are made introduces serious problems for evaluation and trust in humanities research. If we cannot read code or see the workings of the software as it functions, we can experience it only in terms of its interface and its output, neither of which seem subject to our control."

of visualization and discusses those well suited for specific questions and presents considerable dangers to avoided when visualizing. Peters presents a history of information visualization within theological projects. Ultimately, a digital edition can also be the result. But what does this result actually mean? What should a digital edition contain and how does it differ from a digitized edition? **Annette von Stockhausen** clarifies these questions in the final essay in the section.

3.4 Dissemination

As part of the Digital Humanities, Computational Theology not only has to use specific media and forms of analysis but must also engage with new practices of scientific transfer. These forms are presented in a dedicated section. **Ulrike Wuttke** initiates this with a comprehensive introduction to science communication and its significance for DH. Wuttke then examines community building, which also plays an important role in the DH, as its members are forced to build and maintain new networks beyond subject and institute boundaries due to the high degree of interdisciplinarity. The collaborative work that particularly characterizes the DH-Community takes place in proximity to virtual research environments. **Caroline T. Schroeder** describes the advantages and challenges of this relationship. The fact that not only texts are created, but also a wealth of other data can be generated, which should be findable, accessible to everyone, interoperable and reusable, requires well-considered research data management. **Jochen Apel** deals with this crucial aspect. Dissemination practices also include the communication of DH approaches. The use of artificial intelligence in teaching is an obvious example here. **Johanna Gröpler, Margret Mundorf,** and **Nicolaus Wilder** discuss the topic of AI-supported text production in higher education. At this point, it becomes clear that computationally generated products pose new challenges for the entire reviewing process. How can digital editions, text collections, and tools be reasonably evaluated? **Ulrike Henny-Krahmer** discusses this evaluation process in a chapter on reviewing digital resources. All contributions bear witness to a change in conventional research environments and forms of publication. Finally, **Clifford Anderson** reflects on the extent of the changes to the entire academic enterprise brought about by the advent of DH.

4. Acknowledgements

This compendium would not have been possible without the institutional and financial support of a broadly diversified community.

The TheoLab has established itself as a fruitful place of exchange and learning, so our thanks go to all those who have been invited to colloquia, conferences, and lunch talks and have contributed to the growing field of Computational Theology. Stefan Karcher deserves special mention for his initiatives in the context of InFoDiTex and TheoLab – we are also reaping the fruits of his work with this volume. We would like to thank Selina Fucker for her support of the TheoLab, first as an assistant and now as coordinator. We would also like to thank our advisory board, Prof. Helmut Schwier, Prof. Friederike Nüssel, and Prof. Michael Gertz, for their support over the last few years. Prof. Winrich Löhr has also helped us in many ways, for which we would also like to express our gratitude.

We would also like to thank our colleagues from the DH for their friendly acceptance into the networks and discussion forums in which we were able to take our first steps in digital research and its reflection. In particular, we would like to thank the "Digital Humanities Theory" working group of the *Association for Digital Humanities in the German Speaking Areas* (DHd), the DH2023 program committee in Graz, and the DH Lab of the Leibniz Institute of European History in Mainz. We would like to thank our colleagues from the field of theology for their willingness to venture into such an explorative field and to think with us about the development of Computational Theology.

The work on this compendium was financially supported by the FEST Heidelberg, within the framework of the Excellence Strategy of the German federal and state governments through funds from the Research Council of the Field of Focus 3 of Heidelberg University and through the publication fund of the DHd, for which we would also like to express our sincere thanks. The staff of Heidelberg University Press, Dr. Maria Effinger, Anja Konopka, Daniela Jakob and Frank Krabbes, as well as Gunther Gebhard and Steffen Schröter from text plus form were responsible for the form and printing of the volume, for which we owe thanks.

We would especially like to thank our authors: The integration of unfamiliar research practices in such a traditional discipline as theology, with all its existing structures and traditions, requires a high degree of competence in scholarly communication and is a correspondingly great challenge that we could not have faced alone. We would therefore like to thank them for their willingness to open an interdisciplinary dialogue with the authors of the second volume and thus pave them a way into the field of Computational Theology. We would also like to thank the many people from the DH community who have agreed to act as anonymous peer reviewers, to think their way into the contributions and deepen the transfer of knowledge even further. The compendium is thus a practice in what the volume hopes to initiate: Collaborative, transdisciplinary thinking on theological research and teaching with the means of Digital Humanities.

Bibliography

Adler, M., Bartsch, S., Becker, M., Bender, M., Borek, L., Brinkmann, C., Dimpel, F., Fischer, R., Glawion, A., Gülden, S. A., Hastik, C., Hegel, P., Herget, K., Horn, F., Krause, C., Müller, M., Núnez, A., Rapp, A., Scharrer, L., Schmid, O., Stegmeier, J., Thull, B., & Weitin, Th. (2020). *Digitale Philologie: Das Darmstädter Modell.* Darmstadt [= *Digital Philology | Working Papers in Digital Philology,* 1]. DOI: https://doi.org/10.25534/tuprints-00012476 (Accessed: 25 May 2024).

Berry, D., & Fagerjord, A. (2017). *Digital Humanities. Knowledge and Critique in a Digital Age.* Cambridge/Malden: Polity Press.

Burke, A. & Hiepel, L. (2021). Digitale Transformation in der Wissenschaft. Beobachtungen aus Forschung und Lehre. In eid., V. Niggemeier & B. Zimmermann (eds.), *Theologiestudium im digitalen Zeitalter* (pp. 19–24). Suttgart: W. Kohlhammer.

Dalferth, I. (2006). Vorwort. In id. (ed.), *Eine Wissenschaft oder viele? Die Einheit evangelischer Theologie in der Sicht ihrer Disziplinen* (pp. 5 f.). Leipzig: Evangelische Verlagsanstalt [= *Forum Theologische Literaturzeitung,* 17].

Id. (2018). Wirkendes Wort. Bibel, Schrift und Evangelium im Leben der Kirche und im Denken der Theologie. Leipzig: Evangelische Verlagsanstalt.

Dombrowski, Q. (2023). Does Coding Matter for Doing Digital Humanities? In J. O'Sullivan (ed.), *The Bloomsbury Handbook to the Digital Humanities* (pp. 137–145). London/New York/Dublin: Bloomsbury Academic.

Gramelsberger, G. (2023). *Philosophie des Digitalen zur Einführung.* Hamburg: Junius.

Hamidović, D. (2016). An Introduction to Emerging Digital Culture. In C. Clivaz, P. Dilley & D. Hamidović (eds.), *Ancient Worlds in Digital Culture* (pp. 1–16). Leiden/Boston: Brill [= *Digital Biblical Studies* 1].

Heyden, K., & Schröder, B. (2020). Zu diesem Heft. In eid. (eds.), *Theologie im digitalen Raum* (pp. 82–89). Gütersloh: Gütersloher Verlagshaus [= *Verkündigung und Forschung,* 65(2)].

Hutchings, T., & Clivaz, C. (2021). Introduction. In eid. (eds.), Digital Humanities and Christianity. An Introduction (pp. 1–17). Berlin/Boston: De Gruyter [= *Introductions to Digital Humanities. Religion,* 4].

Kirschenbaum, M. (2010). What Is Digital Humanities and What's It Doing in English Departments?, *ADE Bulletin,* 150, 1–7. URL: https://www.uvic.ca/humanities/english/assets/docs/kirschenbaum.pdf (Accessed: 25 May 2024).

Kohle, H. (2018). Geleitwort. In P. Kuroczynski, P. Bell, L. Dieckmann, S. Hoppe & S. Münster (eds.), *Computing Art Reader. Einführung in die digitale Kunstgeschichte* (pp. 15–17). Heidelberg: arthistoricum.net [= *Computing in art and architecture,* 1].

Krämer, S. (2019). Was bedeutet 'Geist'? Eine etwas andere 'Philosophie des Geistes' mit Seitenblick auf die digitalen Geisteswissenschaften. In H. Joas & J. Noller (eds.), *Geisteswissenschaften – was bleibt? Zwischen Theorie, Tradition und*

Transformation (pp. 224–249). Freiburg/München: Karl Alber [= *Geist und Geisteswissenschaft,* 5].
Krautter, B., Pichler, A., & Reiter, N. (2023) Operationalisierung. In AG Digital Humanities Theorie des Verbandes Digital Humanities im deutschsprachigen Raum e.V. (ed.), *Begriffe der Digital Humanities. Ein diskursives Glossar.* Wolfenbüttel [= *Zeitschrift für digitale Geisteswissenschaften. Working Papers,* 2]. DOI: https://doi.org/10.17175/wp_2023_010 (Accessed: 25 May 2024).
Luhmann, J., & Burghardt, M. (2021). Digital Humanities – A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape, *Journal of the Association for Information Science and Technology,* 73(2), 148–171. DOI: https://doi.org/10.1002/asi.24533 (Accessed: 25 May 2024).
Manovich, L. (2020). *Cultural Analytics.* Cambridge, Mass./London: The MIT Press.
Moretti, F. (2000). Conjectures on World Literature, *New Left Review,* 1, 54–68. URL: https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature (Accessed: 25 May 2024).
Nunn, Ch. (2024). Vom Gründungsmythos in die Peripherie… und zurück?! Digital Humanities in der theologischen Forschung, *Kerygma und Dogma,* 70(4), 318–334.
Nüssel, F. (2006). Die Aufgabe der Dogmatik im Zusammenhang der Theologie. In I. Dalferth (ed.), *Eine Wissenschaft oder viele? Die Einheit evangelischer Theologie in der Sicht ihrer Disziplinen* (pp. 77–98). Leipzig: Evangelische Verlagsanstalt [= *Forum Theologische Literaturzeitung,* 17].
van Oorschot, F. (2021), Neue Technik – neue Wissenschaft? Wissenschaftstheoretische und -ethische Herausforderungen der Digital Humanities. In ead. & B. Held (eds.), *Digitalisierung: Neue Technik, neue Ethik? Interdisziplinäre Auseinandersetzung mit den Folgen der digitalen Transformation* (pp. 143–164). Heidelberg: heiUP [= *FESTForschung* 1]. DOI: https://doi.org/10.11588/heibooks.945.c12680 (Accessed: 25 May 2024).
van Oorschot, F. (2023). *Digitale Theologie und digitale Kirche. Eine Orientierung.* Heidelberg: heiBOOKS [= *FEST kompakt,* 7]. DOI: https://doi.org/10.11588/heibooks.1271 (Accessed: 25 May 2024).
Rapp, A. (2021). Digital Humanities und Bibliotheken. Traditionen und Transformationen, *027.7 Zeitschrift für Bibliothekskultur,* 8(1), 1–17. DOI: https://doi.org/10.21428/1bfadeb6.486c17e5 (Accessed: 25 May 2024).
Reiche, R., Becker, R., Bender, M., Munson, M., Schmunk, S., & Schöch, Ch. (2014). Verfahren der Digital Humanities in den Geistes- und Kulturwissenschaften, *DARIAH-DE working papers,* 4, 1–35. URL: http://resolver.sub.uni-goettingen.de/purl/?dariah-2014-2 (Accessed: 25 May 2024).
Ressel, A. (2017). *Basiswissen Digital Humanities (Bd. 1). Grundlagen und Arbeitstechniken.* Göttingen: Cuvillier Verlag.

Sahle, P. (2013). DH studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities, DARIAH-DE working papers, 1, 1–37. URL: http://resolver.sub.uni-goettingen.de/purl/?dariah-2013-1 (Accessed: 25 May 2024).

Id. (2017). Digital Humanities und die Fächer. Eine schwierige Beziehung?, *Forum Exegese und Hochschuldidaktik. Verstehen von Anfang an,* 2(2), 7–28.

Salmi, H. (2021). *What is Digital History?* Cambridge/Medford: Polity Press.

Schelbert, G. (2018). Digital Art History – Digitale Kunstgeschichte. Überlegungen zum aktuellen Stand. In P. Kuroczynski, P. Bell, L. Dieckmann, S. Hoppe & S. Münster (eds.), *Computing Art Reader. Einführung in die digitale Kunstgeschichte* (pp. 41–57). Heidelberg: arthistoricum.net [= *Computing in art and architecture,* 1].

Schmale, W. (2016). Digital Musicology im Kontext der Digital Humanities. In S. Bolz, M. Kelber, I. Knoth & A. Langenbruch (eds.), *Wissenskulturen der Musikwissenschaft. Generationen – Netzwerke – Denkstrukturen* (299–310). Bielefeld: transcript [= *Musik und Klangkultur,* 16].

Schubert, Ch. (2015). Editorial. Close Reading und Distant Reading. Methoden der Altertumswissenschaften in der Gegenwart, *Digital Classics Online,* 1–6. DOI: https://doi.org/10.11588/dco.2015.1 (Accessed: 25 May 2024).

Thaller, M. (2017). Digital Humanities als Wissenschaft. In F. Jannidis, H. Kohle & M. Rehbein (eds.), *Digital Humanities. Eine Einführung* (pp. 13–18). Stuttgart: J.B. Metzler.

Zahnd, U. (2020). Netzwerke, historisch und digital. Digital Humanities und die Mittlere und Neue Kirchengeschichte. In K. Heyden & B. Schröder (eds.), *Theologie im digitalen Raum* (pp. 114–123). Gütersloh: Gütersloher Verlagshaus [= *Verkündigung und Forschung,* 65(2)].

van Zundert, J., Antonijević, S., & Andrews, T. (2020). 'Black Boxes' and True Colour. A Rhetoric of Scholarly Code. In J. Edmond (ed.), *Digital Technology and the Practices of Humanities Research* (pp. 123–162). Cambridge: Open Book Publishers. DOI: https://doi.org/10.11647/OBP.0192 (Accessed: 25 May 2024).

Figure Credit

Fig. 1: Spheres of the Digital Humanities – created by Patrick Sahle, CC BY 4.0.

II. PROJECT AND DEFINITION

Schism or Renaissance?

On the Relationship Between Computational Humanities and Digital Humanities

Michael Piotrowski

https://orcid.org/0000-0003-3307-5386

Abstract The recent emergence of Computational Humanities is often regarded as a secession from Digital Humanities. The narrative of a schism is tempting, but inherently negative and unproductive. Instead, I propose to regard digital and Computational Humanities as heirs of two different traditions. From 2004 onwards, the Anglo-American Digital Humanities became the dominant current and mostly supplanted local European traditions. Although this has certainly been beneficial in some respects, Anglo-American and European academic traditions and institutional contexts differ substantially. The fundamental difference between digital and Computational Humanities is thus less one of *digital* vs. *computational*, but rather one of *humanities* vs. *Geisteswissenschaften:* Digital Humanities are in the former, Computational Humanities in the latter tradition. I therefore reject the notion of a schism and suggest regarding the emergence of Computational Humanities as a sign of renewed interest in the continental tradition of formal methods in the humanities.

Keywords Computational Humanities, Digital Humanities, History, Definition

1. Introduction

Back in 2019, when Twitter[1] was still the town square of Digital Humanities (Grandjean 2016), a tweet by Folgert Karsdorp rocked the community. Shortly after the DH 2019 conference in Amsterdam, Karsdorp posted the following message:

1 As I am writing this (end of July 2023), Elon Musk, who had bought the service in October 2022, has started rebranding Twitter as X. In the light of this and previous changes, the long-term availability of tweets (Twitter posts) is unclear.

Michael Piotrowski: Schism or Renaissance? On the Relationship Between Computational Humanities and Digital Humanities, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 31–49. DOI: https://doi.org/10.11588/heibooks.1521.c21938

> I'm thinking about developing a workshop/event/journal/community for computational research in the humanities that doesn't exclude people with profound computational skills and knowledge. Who's in?[2]

According to Dombrowski (2023, 138), "[t]he response to this turn of events was swift and negative from many parts of DH Twitter,"[3] but this fails to mention that by other parts of the community, Karsdorp's musings were met with "overwhelming enthusiasm and the clear demand for a Computational Humanities research venue,"[4] which prompted Karsdorp, together with other scholars, to proceed further, create a discussion forum,[5] and organize the first Computational Humanities Research workshop, CHR 2020, since then held annually.

Digital Humanities and Computational Humanities, a classic schism, all too familiar to theologians? Heretics that chose to leave the Church of the Big Tent? In the light of the founding myth of Digital Humanities, with Father Busa as the founding father acting on orders he received from God – *"digitus Dei est hic"* (Busa 2004, xvi) – the religious analogy does not seem too far-fetched.

This is a well-known story pattern and thus a compelling narrative for the historicization of the events; Dombrowski (2023, 138) consistently talks of a "splintering-off" of a "group," supposedly homogenous and privileged.[6] However, I do not think that this is the whole story. For a better understanding, we first have to look at the genesis of Digital Humanities.

2 See https://twitter.com/FolgertK/status/1151167545539477504 (Posted: 16 June 2019; accessed: 21 June 2024). Archive link: https://web.archive.org/web/https://twitter.com/FolgertK/status/1151167545539477504 (Accessed: 5 June 2024).

3 For example, Miriam Posner criticized the initiative as "a method of protecting prestige, which tracks closely with masculinization" (see https://twitter.com/miriamkp/status/1152389216363401216, posted: 20 June 2019, accessed: 5 June 2024); in addition to these accusations of elitism and sexism, Roopika Risam alleged racism (see https://twitter.com/roopikarisam/status/1152389797882863617, posted: 20 July 2019, accessed: 5 June 2024), and others heaped further allegations on the initiative in the replies.

4 See https://twitter.com/FolgertK/status/1151572736730439684 (Posted: 17 July 2019, accessed: 5 June 2024. Archive link: https://web.archive.org/web/https://twitter.com/FolgertK/status/1151572736730439684 (Accessed: 5 June 2024).

5 See https://discourse.computational-humanities-research.org (Accessed: 5 June 2024).

6 Similar claims had been made earlier by Lang (2020), who asserted "that they [the Computational Humanities people] are quite over-represented at conferences already. They get all the attention. They get the grants." However, this is not substantiated, and it would be difficult to do so in general. Research funding very much depends on national funding institutions, policies, and politics, so the sweeping claim that "there also are many grants specifically targeted towards CH where DH-only people can't even apply anymore, cutting funding for 'normal' DH projects" is impossible to verify outside a specific academic system.

2. The Genesis of Digital Humanities

Despite the well-known myth of the founding of the field by the Italian Jesuit priest Roberto Busa (Hockey 2004, 4), what we now call "Digital Humanities" is in fact the result of a particular Anglo-American development.[7] For the purposes of our discussion, it suffices to know that the use of digital computers in the humanities in the US dates back to the mid-1950s. Bowles (1967), one of the first books wholly dedicated to the topic, already shows a considerable breadth of applications across many humanities disciplines. At the same time, we also see the beginnings of institutionalization: in 1966, the first issue of the emerging field's first journal, *Computers and the Humanities,* was published. It still bore the modest subtitle "A Newsletter," but in the first article, titled "The Next Step," the author confidently asserts: "We are now moving into the phase of consolidation" (Milic 1966, 3).

In the 1970s, the first learned societies are founded: in 1973, the *Association for Literary and Linguistic Computing* (ALLC) in the UK, and in 1978, its US counterpart, the *Association for Computers and the Humanities* (ACH). The publication of the *Humanities Computing Yearbooks* 1988 and 1989–1990 (Lancashire & McCarty 1989; Lancashire 1991) might be considered the culmination of this work aiming to consolidate the field. However, while reviewers (e.g., Kenner 1989; Potter 1992) found the yearbooks to be useful, they often stressed a lack of coherence: "a *vast* field," but "in a state of free-fall" (Kenner 1989, 360), a "large baggy monster" (Potter 1992, 548). The fact that only two yearbooks were published is thus probably not only due to the size of the field and the rise of the Web (rendering printed directories obsolete), but also to the absence of a unifying vision. On the other hand, the Web probably fostered the adoption of the Text Encoding Initiative's *TEI Guidelines*[8] when they were published in 1994: not only did they put an end to the chaos of incompatible text encoding schemes (Hockey 2004, 12), they also finally brought a consolidated vision, admittedly only for part of the field – mainly philologists – but a very influential one.

The introduction of the term *Digital Humanities,* which eventually replaced earlier terms such as *humanities computing* or *computing in the humanities,* is usually associated with the publication of *A companion to Digital Humanities* (Schreibman et al. 2004a). Even though the volume aims to cover many different humanities disciplines,

7 Digital Humanities has only recently begun to take an interest in its history beyond the "canonical narrative." Publications on the subject are still relatively few, and Sula & Hill (2019, 191) note that despite the variety and breadth of definitions of Digital Humanities (e.g., Gold 2012; Terras et al. 2013), "narratives of its history have been surprisingly homogenous": "all ground DH in mid-twentieth-century humanities computing." The same goes for non-English histories (e.g., Berra 2015; Thaller 2017; Mounier 2018). Burdick et al. (2012) explicitly exclude a discussion of the history of the field; Drucker (2021) only briefly mentions the history of specific methods in the corresponding chapters.

8 See https://tei-c.org (Accessed: 5 June 2024).

the success of the TEI was certainly an impetus for this new attempt at consolidating the field.

As John Unsworth, one of the editors of Schreibman et al. (2004a), related, the editors had originally used "humanities computing" in the title; for marketing reasons, the publisher proposed "digitized humanities," and Unsworth countered with "Digital Humanities" (Kirschenbaum 2012, 5). However, Unsworth did not come up with term spontaneously. In fact, he and colleagues at the University of Virginia had already been using it since at least 2001, in the context of an interdisciplinary seminar "Is Humanities Computing an Academic Discipline?"[9] which also resulted in the proposal for an M.A. in Digital Humanities at the University of Virginia in 2002 [it was not implemented, though; see Rockwell ([1999] 2013), 30]. The "rapid and remarkable rise of *Digital Humanities* as a term" (Kirschenbaum 2012, 56) was certainly also helped by Unsworth's implication in a project aiming to merge the ALLC and the ACH, which ultimately led to the creation of an umbrella organization, the *Association of Digital Humanities Associations* (ADHO) – again, the same people advocating for the same term.

Against this backdrop, Schreibman et al. (2004a) did not aim to create a *new* field of research distinct from humanities computing. The stated aim of the editors – which may come as a surprise – was to envisage humanities computing (or Digital Humanities, the terms are used interchangeably) – as a discipline in its own right rather than making humanities computing "more palatable to humanists in general" (Fitzpatrick 2012, 12).[10] The first paragraph of the introduction reads:

> This collection marks a turning point in the field of digital humanities: for the first time, a wide range of theorists and practitioners, those who have been active in the field for decades, and those recently involved, disciplinary experts, computer scientists, and library and information studies specialists, have been brought together *to consider digital humanities as a discipline in its own right,* as well as to reflect on how it relates to areas of traditional humanities scholarship (Schreibman et al. 2004b, xxiii, emphasis added).

9 See http://www.iath.virginia.edu/hcs (Accessed: 5 June 2024); archive link: https://web.archive.org/web/20010501000000*/http://jefferson.village.virginia.edu/hcs (Accessed: 5 June 2024).

10 This is not to say that the editors and the scholars involved in the University of Virginia seminar were happy with the name *humanities computing.* For example, in an essay originally written for the seminar on the question of disciplinarity at the University of Virginia in November 1999 (see Rockwell [1999] 2013, 32), Rockwell puts forward a number of arguments against "humanities computing" as a name for a study program, e.g., "Humanities Computing is meaningless to people outside its traditions and the program was unlikely to be approved with such an awkward name," or: "Too often humanities computing is focused exclusively on textual computing and is therefore only of interest to students in textual disciplines like English, Linguistics, and Comparative Literature" (Rockwell [1999] 2013, 20).

Thus, despite acknowledging “how broadly the field now defines itself” (Schreibman et al. 2004b, xxiii), further down, the editors stress the continuity: in the conclusion, they talk again about “the digital humanities as they have evolved over the past half century” (Schreibman et al. 2004b, xxvi). The intention to consolidate and institutionalize the field “as a discipline in its own right” must therefore be considered a central concern of the editors.

The actual outcome, however, was almost exactly the opposite: the new term turned out to be so much more “palatable to humanists” that they effectively took over humanities computing under the new name of “Digital Humanities.” This led to an opening towards a huge number of fields and interpretations (“big tent”), which naturally led to a dilution of computing, perhaps best exemplified by the still ongoing debate whether programming skills are necessary in Digital Humanities (see, e.g., Ramsay [2011] 2013; Dombrowski 2023).[11] Instead of the establishment of Digital Humanities as “a discipline in its own right,” the result was a massive *rejection* of discipline formation.

The “big tent” metaphor to emphasize the diversity, openness, and fluidity of Digital Humanities, can be traced as far back as the DH 2011 conference, entitled “Big Tent Digital Humanities.” It is then used to explain (often proudly) that, as a result, Digital Humanities cannot be defined. While not everybody might conclude that “we’re all digital humanists now” (Mullen [2010] 2013, 238), over a decade later O’Sullivan (2023, 1–2) has to admit that “despite all this investment […] there are still people who think DH is putting pictures of books on WordPress sites,” and laments that “[n]obody wants to talk about who’s in and who’s out, because to do so will inevitably involve exclusion.”[12] His conclusion: “Re-engaging with the question, ‘What are the digital humanities?,’ has never been timelier. DH is everywhere, across all continents and cultures, all intellectual communities and research practices” (O’Sullivan 2023, 2).

3. The Globalization of Digital Humanities

“DH is everywhere, across all continents and cultures,” O’Sullivan (2023, 2) writes. However, the history of Digital Humanities generally ignores the use of computers for humanities research outside of the Anglo-American world. In this respect, the founding myth of Digital Humanities may have less to do with the Italian priest than

11 In a literature analysis, Roth (2019) found that what he calls the “digitized humanities” (creation, curation, and use of digitized data sets) dominate clearly – Roth assigns between 73.9 % and 86.6 % of the contributions to this category. He also finds that the majority of the “numerical humanities” contributions (focusing on the development of mathematical frameworks and computer science methods) “essentially had to do with stylometry” (Roth 2019, 12).

12 “Who’s in and who’s out” is a reference to a highly controversial essay of the same name (Ramsay [2011] 2013).

with American IBM. In fact, computers have also been used in humanities research in other parts of the world, notably in Europe,[13] at least since the late 1950s. For example, in the field of textual criticism, to take the domain of Busa, the French Benedictine Jacques Froger experimented with the use of computers for collation in 1960–1961 (Froger 1970, 211). But Froger does not play any role in the Anglo-American tradition of Digital Humanities, even though he published extensively on the use of computers in the humanities (see, e.g., Froger 1965a, 1965b, 1970) and, even more importantly, on related methodological questions that are still relevant (Froger 1968).[14]

By the 1980s, the use of computers in the humanities in France had been so firmly established that Borillo (1985) remarks:

> l'utilisation du “calcul” s'est généralisée, au point qu'il n'y a plus guère de centre de recherche important en sciences humaines qui n'ait son équipe d'informaticiens. De fait, la statistique, l'analyse des données, les systèmes documentaires automatisés, les bases de données… on fait leur entrée dans de nombreux laboratoires. (Borillo 1985, 5)

But this was hardly noticed in the Anglosphere; in his review of Lancashire and McCarty (1989), Breu (1990, 395) notes: “Der Band ist durch ein starkes Übergewicht amerikanischer Arbeiten gekennzeichnet, über das durch den Markt gerechtfertigte Maß hinaus, was die Autoren selbst eingestehen.”

Of course, developments in Europe took place under different circumstances. Driven by competition with the USSR for global supremacy in the Cold War, the US government generously funded computer science and its applications in all areas. In contrast, Western European governments (perhaps with the exception of France) failed to recognize the strategic importance of computer science and consequently invested little (for a contemporary critique, see, e.g., Steinbuch [1966] 1969). In the Eastern Bloc, the development of computer science was slowed first by Stalinist rejection of cybernetics as bourgeois pseudoscience (see Gerovitch 2002; Shilov 2014) and then by mismanagement and persistent material shortages.

What is more important in the context of this chapter, though, is that the continental European conception of the humanities is quite different from the Anglo-American one, which also leads to a different relationship to computing. There is a long European tradition of formalization in the humanities – for example, Russian formalism, structuralism, the Prague School, the ideas about the unity of science of the positivists or the Vienna Circle, modern hermeneutics (especially Dilthey), Marxism, and so on. When these earlier ideas were combined with the then new ideas of cybernetics and information theory from the 1950s onwards, they all provided moti-

13 I will limit my discussion to Europe.

14 The first, theoretical, part of this book has been recently republished with commentary (Poirel [1968] 2022).

vations and epistemological frameworks for the use of computers in the humanities that went *beyond* the automation of tedious tasks.

This includes work on epistemology (e.g., Granger [1960] 1967; Klaus 1966), on aesthetics (in particular by Max Bense and Abraham Moles, see, e.g., Bašičević & Picelj 1968), in history (e.g., Topolski [1973] 1976; Le Roy Ladurie 1968; Bautier 1977), in linguistics (e.g., Ceccato 1964), and in archeology (Gardin & Garelli 1961). Even the Austrian computing pioneer Heinz Zemanek explicitly relates information technology to the work of Wittgenstein; his lecture series *Das geistige Umfeld der Informationstechnik* (Zemanek 1992) includes a chapter titled "Computer für die Geisteswissenschaften, Geisteswissenschaften für den Computer," which highlights the two-way connection between computing and the humanities.

Ironically, Busa himself is rooted in these traditions and, of course, in the strong scientific tradition of the Society of Jesus, which has been called "[t]he single most important contributor to the support of the study of experimental physics in the seventeenth century" (Heilbron 1979, 2).

As far as I know, there is no research on when exactly scholars in continental Europe became aware of Digital Humanities, either as a term or a concept. The *THATCamp Paris 2010,* the "first unconference on Digital Humanities in France" is likely to have played a role in popularizing it in continental Europe. The fact that the original French version of the *Manifeste des Digital Humanities* (Mounier 2010) ("Manifesto for the Digital Humanities"[15]) published at this meeting gives the gloss "humanités numériques," but uses only the English term, indicates that the French translation had not yet established itself at that time, and that it was perceived as something *new.* The manifesto notes that "experiments in the digital domain of the social sciences and humanities have multiplied in the last half century," its impassioned language nevertheless signals a new beginning rather than a simple announcement of a new name for an existing field. While the authors proclaim that "digital humanities are not *tabula rasa,*" the continuity applies to the disciplines of the humanities and social sciences, *not* to *informatique pour les sciences humaines* (humanities computing), which is conspicuously absent from the *Manifesto.*

Thus, Digital Humanities, as a global phenomenon, is the product of a distinctly Anglo-American tradition of the humanities, of computing, and of computing in the humanities. It evolved in particular institutional contexts and around certain social practices.

Svensson ([2009] 2013, 174) raises the question as to whether the discursive transition from the term *humanities computing* to the designation *Digital Humanities* was essentially a simple "repackaging" of the former, or whether the new name indicates more fundamental changes, such as a broadening of the field or a shift in focus. The

15 See https://oep.hypotheses.org/78 (Accessed: 5 June 2024) for the English-language version. The quotes in English are from this version.

Manifesto is a potential piece of evidence[16] that when Digital Humanities arrived in continental Europe, it supplanted, rather than invigorated, local traditions of computing in the humanities (e.g., *informatique pour les sciences humaines, geisteswissenschaftliche Fachinformatik* or *informatica umanistica*). It seems that here, Digital Humanities was considered as a completely new field, inspired by Anglo-American models, rather than just a new name. Digital Humanities still tends to be more "computational" in continental Europe than in the US, which is likely due to a different conception (and institutionalization) of the humanities rather than an influence of these local predecessors; along with the names, much of the approaches, practices, and traditions were marginalized or even lost in the transition.

4. A Short History of the Term *Computational Humanities*

The term *Computational Humanities* is not new; in fact, it is older than the term *Digital Humanities.* The first use I have found so far is in an article on future directions of computing. The author gives two examples to demonstrate that "technology is transforming the scholarship" of humanities researchers, not only because it is more convenient, but because "the representation of and access to information allows them to organize kinds and quantities of information that weren't possible, hence to ask and answer questions about the human record that couldn't be answered before" (Wulf 1997, 111). He draws a parallel to science and engineering, where "we are used to the notion that new instruments allow us to address new questions; now the same is happening in the humanities. And just as in the sciences, the enhanced ability to answer questions provokes us to ask questions we hadn't considered before" (Wulf 1997, 111).

The next published occurrence I could find is by Cruz-Neira (2003), who even gives an explicit definition of Computational Humanities as "an emerging field that bridges the sciences and humanities with the goal of creating accurate computer simulations of historical, social, cultural, and religious events" (Cruz-Neira 2003, 10). This seems to be one of the earliest uses of *Computational Humanities* as a fixed term describing an identifiable field.

To my current knowledge, Bock et al. (2013) is the first publication in which *Computational Humanities* is used explicitly to denote a field *distinct* from Digital Humanities:[17]

16 Among others, such as the creation of Digital Humanities programs *alongside* existing programs in humanities computing.

17 The preface in which the editors explain their choice is is actually dated November 2011.

> *Computational humanities* are an emerging discipline, following concepts like the computational sciences in other fields […]. The term *computational* is chosen instead of *digital,* used in the name *digital humanities,* since the spectrum of concepts and methods applied is broader and not focused mainly on information sciences (Bock, Jäger, and Winckler 2013, v, emphasis in original).

The editors stress their "hope to attract the interest especially of young researchers to this young discipline" (Bock et al. 2013, vii) and conclude: "Research in *computational humanities* is a challenge, offering many perspectives" (Bock et al. 2013, vii, emphasis in original).

Around the same time, Zundert et al. (2012) hint at "current efforts at developing computational humanities" (Zundert et al. 2012, 298). In fact, the short bio of Zundert in the volume describes him as "a researcher and developer in the field of computational humanities," a description he had already been using for several years at this point. Even though this contribution does not explicitly define Computational Humanities, Zundert et al. (2012) present a clear vision that goes beyond the use of computational methods as mere tools. Specifically, they highlight the role of formalization as "an integral part of humanities practice and not as a feature driven only by computation" (2012, 287). The authors argue that "the ongoing computational 'waves' and 'turns' should not steer the research community away from maintaining and promoting the traditions of humanities in contemporary scholarship," and that Computational Humanities "should be unequivocally recognised as only one stream of contemporary humanities research" (2012, 288). Even though the volume in which their chapter appears (Berry 2012), titled *Understanding Digital Humanities,* Zundert et al. (2012) does not distinguish Computational from Digital Humanities; in fact, they do not use the latter term at all.

In contrast, Biemann et al. (2014), in their report on the 2014 Dagstuhl seminar "Computational Humanities – Bridging the Gap Between Computer Science and Digital Humanities," intentionally use *Computational Humanities* in contrast to *Digital Humanities* – along with a reflection on its definition and its relationship to the humanities, computer science, and DH. The introduction to the report outlines the organizers' understanding of Computational Humanities (CH) as follows:

> At the core of the organizers' understanding of CH is the idea that CH is a discipline that should provide an algorithmic foundation as a bridge between computer science and the humanities. As a new discipline, CH is explicitly concerned with research questions from the humanities that can more successfully be solved by means of computing. CH is also concerned with pertinent research questions from computing science focusing on multimedia content, uncertainties of digitisation, language use across long time spans and visual presentation of content and form (Biemann et al. 2014, 81).

The organizers thus regard CH as a "new discipline," an "independent field of research," and notably one *distinct from DH.* In his talk (for the abstract, see Biemann et al. 2014, 87), Gerhard Heyer from the University of Leipzig detailed this conception by describing CH and DH as being part of computer science and the humanities, respectively, and *jointly* constituting an interface between computer science and the humanities. Arguing that the "degree of mutual understanding of research issues, technical feasibility and scientific relevance of research results will be much higher in the area of overlap between the Computational and Digital Humanities than with any intersection between Computer Science and the Humanities," he proposed to "set up research groups in both scientific communities, Computer Science and Humanities."[18]

Heyer describes Digital Humanities as "the creation, dissemination, and use of digital repositories" and Computational Humanities as "the computer based analysis of digital repositories using advanced computational and algorithmic methods," which "implies a dominance of computational aspects." Heyer further argues that the difference between the two orientations is reflected "in the know-how of researchers and their organizational attachment to either Humanities or Computer Science departments," and that consequently their research is either "more focused on just the creation and use of digital repositories, or on real program development in the Humanities as an area of applied Computer Science" (Biemann et al. 2014, 87–88). This conception seems similar to that of Bock et al. (2013) cited above.

While the descriptions of Computational Humanities cited above vary somewhat, they largely overlap. I only want to highlight two aspects here. First, they distinguish between digital representations and computational operations; Meunier (2014, 22) remarks: "pour ce programme de recherche, la caractéristique d'être numérique est secondaire. La plus importante est la computationnalité." Consequently, the challenge does not consist in digitizing the artifacts studied by the humanities, but rather to identify their *tasks* and to translate these into computational functions. Second, none of the authors who give a description of Computational Humanities hesitate to call it a *discipline.*

18 This is exactly the configuration that was realized at the University of Leipzig: Digital Humanities and Computational Humanities are two research groups, the former leaning more to the humanities side (now defunct) and the latter more to the computer science side. While relatively rare elsewhere, similar ideas have been used by other institutions; for example, Crum et al. (2019, 389) describe "synchronized courses of computational humanities and digital humanities."

5. Defining Computational Humanities

I have previously proposed an explication (or a stipulative definition) of Digital Humanities (Piotrowski 2018; Piotrowski & Xanthos 2020). I chose to use the term *Digital Humanities* rather than *Computational Humanities* for practical reasons: it was (and still is) the more established term. Since then, I have realized that any attempt at defining Digital Humanities will be met with disinterest. My mistake, however, was not so much in underestimating the resistance to defining Digital Humanities, but in believing that what I was trying to define was a subset of Digital Humanities.

I am now convinced that what I am trying to define should not be understood as a subfield of Digital Humanities, but as a field in its own right, which is part of a different tradition and draws on a long history of formal approaches to the humanities,[19] even if it is clearly linked to certain orientations of Digital Humanities.

My definition is based on the following considerations. First of all, a coherent field of research (which may or may not actually be considered a discipline) is characterized by a particular combination of (1) a research object and (2) a research objective; it is *not* dependent on the use of particular methods. Second, as Granger points out, the goal of any systematic pursuit of knowledge is the "construction de *modèles cohérents et efficaces du phénomène*" (Granger [1960] 1967, 215, emphasis in original). All research builds models, since the study of an object is nothing other than the creation of its model. Stachowiak (1973) affirms: "Hiernach *ist alle Erkenntnis Erkenntnis in Modellen* oder *durch Modelle*, und jegliche menschliche Weltbegegnung überhaupt bedarf des Mediums 'Modell'." (Stachowiak 1973, 56, emphasis in original). Bachelard (1979, 3) characterizes the model as "un intermédiaire à qui nous déléguons la fonction de connaissance, plus précisément de réduction de l'encore-énigmatique, en présence d'un champ d'études dont l'accès, pour des raisons diverses, nous est difficile." In short, we model to understand (Le Moigne 2003).

The importance of the computer lies precisely in the fact that is a "universal modeling machine"; "computers came into existence for the sake of modeling" (Mahoney 2000). More specifically, the massive impact of computers on research is due to that fact that they, as Weizenbaum ([1976] 1984, 144) puts it, "make possible an entirely new relationship between theories and models": "A theory written in the form of a computer program is thus both a theory and, when placed on a computer and run, a model to which the theory applies" (Weizenbaum [1976] 1984, 145).

Thus, the difference between Computational Humanities and most traditional research in the humanities is *not* that Computational Humanities constructs models, but that Computational Humanities constructs *formal* models that can be manipulated by the computer, i.e., *computational models.* This general notion is in line with

19 Perhaps we should (like, e.g., Mazlish [1998] 2017) rather speak of the *human sciences* in this context, since it is essential to start from a conceptualization more akin to the *Geisteswissenschaften* or the *sciences humaines,* rather the Anglo-American humanities.

important earlier, mostly European, work on computing in the humanities. However, we need to further distinguish between two subfields of Computational Humanities, because they clearly differ in their research objectives; I call these subfields *applied Computational Humanities* and *theoretical Computational Humanities.*[20]

Applied Computational Humanities refer to those fields of research which, like computational history or computational literary studies, fall within a humanities discipline and have as their object the construction of formal models of the phenomena studied by this "parent discipline," as well as the methodology of this construction. The difference between the "traditional" and "computational" therefore relates specifically to the nature of the models they aim to construct: in the case of the latter, they are *formal* models that can be manipulated by computers. In all other respects, they share the research objects and objectives of the humanities disciplines to which they belong. In particular, computational research must meet all the quality and relevance criteria of these disciplines – it goes without saying that no relaxation of these criteria can be justified by the use of particular methods and tools.

Theoretical Computational Humanities, on the other hand, studies the general properties of such models at a higher level of abstraction. In other words, theoretical Computational Humanities create and study the *metamodels* whose concrete application in the humanities is the domain of applied Computational Humanities, as well as the methodology for constructing these metamodels. One might say that theoretical Computational Humanities deals with the general theory of materials and construction, while applied Computational Humanities raises the building. Because of their metascientific nature, theoretical Computational Humanities is neither "quantitative" nor "qualitative." The goal of theoretical Computational Humanities is to develop abstract models, metamodels, which may or may not have a quantitative dimension, but the research question underlying them is the *adequacy* of these models, not their application. Theoretical Computational Humanities therefore serves as metascience for applied Computational Humanities.

This distinction is crucial, as applied and theoretical Computational Humanities have different research objects and research objectives: in the former, they belong to the humanities, in the latter, to computer science.

As such, theoretical Computational Humanities could be likened to disciplines like business informatics that integrate aspects from both the application domain and computer science.[21] As Wedekind et al. (1998, 265) point out, "Eine der grundlegenden

20 The original inspiration for what follows comes from the definition for *mathematical linguistics* proposed by Gladkij & Mel'čuk (1969).

21 The discipline of business informatics was first established in Germany, and it is one of the disciplines that are commonly grouped under the heading *angewandte Informatik.* While this literally translates to "applied computer science," it refers to something very different, precisely because the disciplines of *angewandte Informatik* integrate aspects from both the application domain and computer science.

Aufgaben der Informatik besteht darin, die aus den Fachwissenschaften stammenden Modelle so umzuschreiben, daß sie mit Hilfe eines Computers dargestellt und bearbeitet werden können." If such models cannot be translated directly into the language of computing, then the model must be reconstructed. Görz (2018, 164) notes that "[i]n the humanities in particular, understanding and explanation of actions in terms of reasons and intentionality provide challenges to operationalized representations." In other words, an in-depth understanding of both domains is essential, rather than just interdisciplinary mediation, which suggests considering theoretical Computational Humanities as a discipline in its own right.

That said, it is important to stress that there are no strict boundaries between applied and theoretical Computational Humanities: researchers in applied Computational Humanities will often be interested in the general properties of the models they are constructing, whereas researchers in theoretical Computational Humanities will naturally be interested in concrete applications of the metamodels they are developing. And of course, Computational Humanities as a whole is in constant exchange with Digital Humanities, the humanities disciplines, as well as computer science.

As Granger ([1960] 1967, 19) points out: "Ce sera l'un des aspects de notre tâche que de montrer la pensée formelle à l'œuvre dans les sciences humaines, non pas seulement comme réduction des phénomènes aux calculs, mais aussi comme invention de structures nouvelles, voire même d'une mathématique originale." The development of this "original mathematics" – today Granger would probably write "informatique originale" – is, I believe, indeed the task of Computational Humanities.

6. Conclusion

In this chapter, I have proposed a new reading of the genesis of Computational Humanities, as well as a concise definition. While the narrative of a schism may be tempting, it is inherently negative and not very productive. Instead, I propose to regard Digital and Computational Humanities as heirs of two different, although obviously related, traditions – perhaps comparable, in their relation, to analytic and continental philosophy. From 2004 onwards, the Anglo-American Digital Humanities established their global dominance and mostly supplanted local European traditions.

Rallying behind the new term "Digital Humanities" has certainly been beneficial, especially for the legitimization and institutionalization of computer-aided research in the humanities in European universities. However, Anglo-American and European academic traditions and institutional contexts differ substantially. Certain interpretations of Digital Humanities adapt quite well to European contexts, but this is much less true for others. At the global ADHO DH conferences the different conceptions of Digital Humanities become apparent and raise questions, for example, about the adequate evaluation of research, a question that is not limited to the acceptance of papers at the conference, but also touches on issues of recognition and evaluation "at home," in the respective academic systems in which researchers work.

The ideal (one is tempted to say: dogma) of the "big tent" is also strongly rooted in US traditions: it is commonly interpreted as referring to the inclusivity of Digital Humanities; however, it also has a flip side: Digital Humanities as "one field indivisible." Under this ideal, all discussion about evaluation and institutionalization are effectively rendered taboo, because fixed evaluation criteria would either exclude some people, or divide the field.

While such factors have certainly contributed to the frustration of some scholars, the fundamental problem is perhaps less one of *digital* vs. *computational,* but rather one of *humanities* vs. *Geisteswissenschaften:* Digital Humanities are in the former, Computational Humanities in the latter tradition. This is why I reject the notion of a schism and prefer to talk about a re-emergence or renaissance: these different conceptualizations have coexisted for a long time; what we see now is a renewed interest in the continental tradition of formal methods in the humanities – now referred to as "Computational Humanities" – rather than a schism of Digital Humanities.

Acknowledgements

This chapter was written during a stay at the Fondazione Garbald's Studio Cascina in August 2023. I thank Barbara Hof and the anonymous reviewer for their useful comments on the first version of this chapter.

References

Bachelard, S. (1979). Quelques aspects historiques des notion de modèle et de justification des modèles. In P. Delattre & M. Thellier (Eds.), *Élaboration et justification des modelès* (pp. 3–19). Paris: Maloine.

Bašičević, D., & Picelj, I. (Eds.). (1968). The Theory of Information and the New Aesthetics, *Bit,* 1. URL: https://monoskop.org/images/b/bf/Bit_International_1_The_Theory_of_Informations_and_the_New_Aesthetics_1968.pdf (Accessed: 5 June 2024).

Bautier, R.-H. (1977). Les demandes des historiens à l'informatique. La forme diplomatique et le contenu juridique des actes. In L. Fossier, A. Vauchez & C. Violante (Eds.), *Informatique et histoire médiévale. Actes du colloque de Rome (20–22 mai 1975)* (pp. 179–186). Rome, Italy: École Française de Rome [= *Publications de l'École française de Rome* 31]. URL: https://persee.fr/doc/efr_0000-0000_1977_act_31_1_2252 (Accessed: 5 June 2024).

Berra, A. (2015). Pour une histoire des humanités numériques, *Critique,* 8–9(819–820), 613–626. DOI: https://doi.org/10.3917/criti.819.0613 (Accessed: 5 June 2024).

Berry, D.M. (Ed.). (2012). *Understanding Digital Humanities.* Basingstoke: Palgrave Macmillan. DOI: https://doi.org/10.1057/9780230371934 (Accessed: 5 June 2024).

Biemann, Ch., Crane, G.R., Fellbaum, Ch.D., & Mehler, A. (Eds.). (2014). Computational Humanities. Bridging the Gap Between Computer Science and Digital Humanities (Dagstuhl Seminar 14301), *Dagstuhl Reports,* 4(7). Schloss Dagstuhl: Leibniz-Zentrum für Informatik. DOI: https://doi.org/doi.org/10.4230/DAGREP.4.7.80 (Accessed: 5 June 2024).

Bock, H.G., Jäger, W., & Winckler, M.J. (Eds.). (2013). *Scientific Computing and Cultural Heritage.* Berlin/Heidelberg: Springer [= *Contributions in Mathematical and Computational Sciences,* 3]. DOI: https://doi.org/10.1007/978-3-642-28021-4 (Accessed: 5 June 2024).

Borillo, M. (1985). *Informatique pour les sciences de l'homme. Limites de la formalisation du raisonnement.* Brussels: Mardaga.

Bowles, E.A. (Ed.). (1967). *Computers in Humanistic Research. Readings and Perspectives.* Englewood Cliffs: Prentice-Hall.

Breu, W. (1990). Book Reviews. The Humanities Computing Yearbook, *System,* 18(3), 394–396. DOI: https://doi.org/10.1016/0346-251x(90)90016-x (Accessed: 5 June 2024).

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). *Digital_Humanities.* Cambridge: MIT Press.

Busa, R.A. (2004). Foreword. Perspectives on the Digital Humanities." In S. Schreibman, R. Siemens & J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. xvi–xxi). Oxford: Blackwell. DOI: https://doi.org/10.1002/9780470999875.fmatter (Accessed: 5 June 2024).

Ceccato, S. (1964). Automatic Translation of Languages, *Information Storage and Retrieval,* 2(3), 105–158. DOI: https://doi.org/10.1016/0020-0271(64)90002-6 (Accessed: 5 June 2024).

Crum, W.B., Angello, A., Liu, X., & Campion, C. (2019). Enabling Interdisciplinary Instruction in Computer Science and Humanities. An Innovative Teaching and Learning Model Customized for Small Liberal Arts Colleges. In *Proceedings of the ICCS 2019* (pp. 389–400). Cham: Springer. DOI: https://doi.org/10.1007/978-3-030-22750-0_31 (Accessed: 5 June 2024).

Cruz-Neira, C. (2003). Computational Humanities. The New Challenge for VR, *IEEE Computer Graphics and Applications,* 23(3), 10–13. DOI: https://doi.org/10.1109/mcg.2003.1198257 (Accessed: 5 June 2024).

Dombrowski, Q. (2023). Does Coding Matter for Doing Digital Humanities? In J. O'Sullivan (Ed.), *The Bloomsbury Handbook to the Digital Humanities* (pp. 137–145). London: Bloomsbury.

Drucker, J. (2021). *The Digital Humanities Coursebook. An Introduction to Digital Methods for Research and Scholarship.* London/New York: Routledge. DOI: https://doi.org/10.4324/9781003106531 (Accessed: 5 June 2024).

Fitzpatrick, K. (2012). The Humanities, Done Digitally. In M.K. Gold (Ed.), *Debates in the Digital Humanities* (pp. 12–15). Minneapolis: University of Minnesota Press. DOI: https://doi.org/10.5749/minnesota/9780816677948.001.0001 (Accessed: 5 June 2024).

Froger, J. (1965a). La machine électronique au service des sciences humaines, *Diogène,* 52, 110–144.

Id. (1965b). The Electronic Machine at the Service of Humanistic Studies. Translated by Nora McKeon, *Diogenes,* 13(52), 104–142. DOI: https://doi.org/10.1177/039219216501305206 (Accessed: 5 June 2024).

Id. (1968). *La critique des textes et son automatisation.* Paris: Dunod.

Id. (1970). La critique des textes et l'ordinateur, *Vigiliae Christianae,* 24(3), 210–217. DOI: https://doi.org/10.2307/1583073 (Accessed: 5 June 2024).

Gardin, J.-C., & Garelli, P. (1961). Étude des établissements assyriens en Cappadoce par ordinateurs, *Annales. Histoire. Sciences Sociales,* 16(5), 837–876. DOI: https://doi.org/10.3406/ahess.1961.420758 (Accessed: 5 June 2024).

Gerovitch, S. (2002). *From Newspeak to Cyberspeak. A History of Soviet Cybernetics.* Cambridge: MIT Press. DOI: https://doi.org/10.7551/mitpress/3137.001.0001 (Accessed: 5 June 2024).

Gladkij, A.V., & Mel'čuk, I.A. (1969). *Elementy matematičeskoj lingvistiki.* Moscow: Nauka.

Gold, M.K. (Ed.). 2012. *Debates in the Digital Humanities.* Minneapolis: University of Minnesota Press. DOI: https://doi.org/10.5749/minnesota/9780816677948.001.0001 (Accessed: 5 June 2024).

Görz, G. (2018). Some Remarks on Modelling from a Computer Science Perspective, *Historical Social Research. Supplement,* 31, 163–169. DOI: https://doi.org/10.12759/HSR.SUPPL.31.2018.163-169 (Accessed: 5 June 2024).

Grandjean, M. (2016). A Social Network Analysis of Twitter. Mapping the Digital Humanities Community, *Cogent Arts & Humanities,* 3(1). DOI: https://doi.org/10.1080/23311983.2016.1171458 (Accessed: 5 June 2024).

Granger, G.-G. (1960 [1967]). *Pensée formelle et sciences de l'homme.* Nouvelle éd. augmentée d'une préface. Paris: Aubier-Montaigne.

Heilbron, J.L. (1979). *Electricity in the 17th and 18th Centuries. A Study of Early Modern Physics.* Berkely/Los Angeles/London: University of California Press.

Hockey, S. (2004). The History of Humanities Computing. In S. Schreibman, R. Siemens & J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 1–19). Oxford: Blackwell. DOI: https://doi.org/10.1002/9780470999875.ch1 (Accessed: 5 June 2024).

Kenner, H. (1989). Curtains for Lady Buxley, *Byte,* 14, 360–361.

Kirschenbaum, M. G. (2012). What Is Digital Humanities and What's It Doing in English Departments? In M. K. Gold (Ed.), *Debates in the Digital Humanities* (pp. 3–11). Minneapolis: University of Minnesota Press. URL: https://www.jstor.org/stable/10.5749/j.ctttv8hq.4 (Accessed: 5 June 2024).

Klaus, G. (1966). *Kybernetik und Erkenntnistheorie.* Berlin: VEB Deutscher Verlag der Wissenschaften.

Lancashire, I., (Ed.). (1991). *The Humanities Computing Yearbook 1989–90. A Comprehensive Guide to Software and Other Resources.* Oxford: Clarendon Press.

Lancashire, I., & McCarty, W. (Eds.). (1989). *The Humanities Computing Yearbook 1988.* Oxford: Clarendon Press.

Lang, S. (2020). News on the DH and Gender Equality, *LaTeX Ninja blog.* URL: https://latex-ninja.com/2020/05/24/news-on-the-dh-and-gender-equality (Accessed: 5 June 2024).

Le Moigne, J.-L. (2003). *Le constructivisme. Modéliser pour comprendre,* 3. Paris: L'Harmattan.

Le Roy Ladurie, E. (1968). La fin des érudits, *Le Nouvel Observateur,* 182, 38–39.

Mahoney, M. S. (2000). Historical Perspectives on Models and Modeling. In *XIIIth DHS-DLMPS Joint Conference on Scientific Models. Their Historical and Philosophical Relevance.* Zürich. URL: https://www.princeton.edu/~hos/Mahoney/articles/models/models.html (Accessed: 5 June 2024).

Mazlish, B. (1998 [2017]). *The Uncertain Sciences. With a New Introduction by the Author.* London/New York: Routledge.

Meunier, J.-G. (2014). Humanités numériques ou computationnelles. Enjeux herméneutiques, *Sens public.* URL: http://sens-public.org/articles/1121 (Accessed: 5 June 2024).

Milic, L. T. (1966). The Next Step, *Computers and the Humanities,* 1(1), 3–6. DOI: https://doi.org/10.1007/bf00188010 (Accessed: 5 June 2024).

Mounier, P. (2010). Manifeste des Digital Humanities, *Journal des anthropologues,* 122–123, 447–452. DOI: https://doi.org/10.4000/jda.3652 (Accessed: 5 June 2024).

Id. (2018). *Les humanités numériques. Une histoire critique.* Paris: Éditions de la Maison des sciences de l'homme. DOI: https://doi.org/10.4000/books.editionsmsh.12006 (Accessed: 5 June 2024).

Mullen, L. A. (2010 [2013]). Digital Humanities Is a Spectrum, or 'We're All Digital Humanists Now'. In M. Terras, J. Nyhan & E. Vanhoutte (Eds.), *Defining Digital Humanities* (pp. 237–238). Farnham: Ashgate. DOI: https://doi.org/10.4324/9781315576251 (Accessed: 5 June 2024).

O'Sullivan, J. (2023). Introduction. Reconsidering the Present and Future of the Digital Humanities. In id. (Ed.), *The Bloomsbury Handbook to the Digital Humanities* (pp. 1–4). London: Bloomsbury.

Piotrowski, M. (Ed.). (2018). *Proceedings of the Workshop on Computational Methods in the Humanities.* Lausanne: COMHUM 2018. URL: http://ceur-ws.org/Vol-2314 (Accessed: 5 June 2024).

Piotrowski, M., & Xanthos, A. (2020). Décomposer les humanités numériques, *Humanités numériques,* 1, 1–35. DOI: https://doi.org/10.4000/revuehn.381 (Accessed: 5 June 2024).

Poirel, D. (1968 [2022]). *Stemma Codicum. La “Méthode Froger” d’édition Critique.* Turnhout: Brepols.

Potter, R.G. (1992). Book Reviews. The Humanities Computing Yearbook 1989–90, *Computational Linguistics,* 18(4), 545–548. URL: https://aclanthology.org/J92-4008 (Accessed: 5 June 2024).

Ramsay, S. (2011 [2013]). Who’s in and Who’s Out. In M. Terras, J. Nyhan & E. Vanhoutte (Eds.), *Defining Digital Humanities* (pp. 239–241). Farnham: Ashgate. DOI: https://doi.org/10.4324/9781315576251 (Accessed: 5 June 2024).

Rockwell, G. (1999 [2013]). Is Humanities Computing an Academic Discipline. In M. Terras, J. Nyhan & E. Vanhoutte (Eds.), *Defining Digital Humanities* (pp. 13–13). Farnham: Ashgate. DOI: https://doi.org/10.4324/9781315576251 (Accessed: 5 June 2024).

Roth, C. (2019). Digital, Digitized, and Numerical Humanities, *Digital Scholarship in the Humanities,* 34(3), 616–632. DOI: https://doi.org/10.1093/llc/fqy057 (Accessed: 5 June 2024).

Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2004a). *A Companion to Digital Humanities.* Oxford: Blackwell. DOI: https://doi.org/10.1002/9780470999875 (Accessed: 5 June 2024).

Eid. (2004b). The Digital Humanities and Humanities Computing. An Introduction. In eid. (Eds.), *A Companion to Digital Humanities* (pp. xxii–xxvii). Oxford: Blackwell. DOI: https://doi.org/10.1002/9780470999875.fmatter (Accessed: 5 June 2024).

Shilov, V. (2014). Reefs of Myths. Towards the History of Cybernetics in the Soviet Union. In *Proceedings of the Third International Conference on Computer Technology in Russia and in the Former Soviet Union 2014.* Washington, DC: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.1109/sorucom.2014.46 (Accessed: 5 June 2024).

Stachowiak, H. (1973). *Allgemeine Modelltheorie.* Wien/New York: Springer.

Steinbuch, K. (1966 [1969]). *Die informierte Gesellschaft. Geschichte und Zukunft der Nachrichtentechnik.* Newly revised, updated Edition. Stuttgart: Deutsche Verlags-Anstalt.

Sula, Ch.A., & Hill, H.V. (2019). The Early History of Digital Humanities. An Analysis of Computers and the Humanities (1966–2004) and Literary and Linguistic Computing (1986–2004), *Digital Scholarship in the Humanities,* 34, 190–206. DOI: https://doi.org/10.1093/llc/fqz072 (Accessed: 5 June 2024).

Svensson, P. (2009 [2013]). Humanities Computing as Digital Humanities. In M. Terras, J. Nyhan & E. Vanhoutte (Eds.), *Defining Digital Humanities* (pp. 159–186). Farnham: Ashgate. DOI: https://doi.org/10.4324/9781315576251 (Accessed: 5 June 2024).

Terras, M., Nyhan, J., & Vanhoutte, E. (Eds.). (2013). *Defining Digital Humanities.* Farnham: Ashgate. DOI: https://doi.org/10.4324/9781315576251 (Accessed: 5 June 2024).

Thaller, M. (2017). Geschichte der Digital Humanities. In F. Jannidis, H. Kohle & M. Rehbein (Eds.), *Digital Humanities. Eine Einführung* (pp. 3–12). Stuttgart: J.B. Metzler.

Topolski, J. (1973[1976]). *Methodology of History.* Translated by O. Wojtasiewicz. Dordrecht: Reidel [= *Synthese Library* 88]. DOI: https://doi.org/10.1007/978-94-010-1123-5 (Accessed: 5 June 2024).

Wedekind, H., Görz, G., Kötter, R., & Inhetveen, R. (1998). Modellierung, Simulation, Visualisierung. Zu aktuellen Aufgaben der Informatik, *Informatik-Spektrum,* 21(5), 265–272. DOI: https://doi.org/10.1007/s002870050104 (Accessed: 5 June 2024).

Weizenbaum, J. (1976 [1984]). *Computer Power and Human Reason. From Judgement to Calculation.* With new preface. Harmondsworth: Penguin.

Wulf, W.A. (1997). Look in the Spaces for Tomorrow's Innovations, *Communications of the ACM,* 40(2), 109–111. DOI: https://doi.org/10.1145/253671.253744 (Accessed: 5 June 2024).

Zemanek, H. (1992). *Das geistige Umfeld der Informationstechnik.* Berlin/Heidelberg: Springer.

Zundert, J.v., Antonijević, S., Beaulieu, A., Dalen-Oskam, K.v., Zeldenrust, D. & Andrews, T.L. (2012). Cultures of Formalisation. Towards an Encounter Between Humanities and Computing. In D.M. Berry (Ed.), *Understanding Digital Humanities* (pp. 279–294). Basingstoke: Palgrave Macmillan. DOI: https://doi.org/10.1057/97802303719 (Accessed: 5 June 2024).

Doing Theology *with* Videogames – Insights for Computational Theology

Erin Raffety

https://orcid.org/0000-0002-6543-6513

Abstract This chapter explores the intersections of Computational Humanities, Digital Theology, and videogames, asserting that a definition of Computational Theology must clarify how its assorted methods, both digital and computational, inform knowledge-making, and avoid separating users from computers, something the field of *Human Computer Interaction* (HCI) has been instrumental in establishing (especially when it comes to disabled users). Drawing on a mixed-methods pilot study, in which disabled users helped to design and play a videogame to enrich Protestant Christian faith experiences in the Northeastern United States, the author provides three guiding insights for doing Computational Theology with videogames:
(1) Computational theologians wishing to engage videogames must center users in prototype development and methods for study;
(2) Computational theologians must appreciate play as a site of theological knowledge-making, moving from observing structural symmetry between religion and games to doing theology with games and gamers themselves;
(3) Even when games are created with specific users in mind, computational theologians must not mistake games for neutral objects; rather, they must continually interrogate the theological underpinnings of computational models.

Keywords Computational Theology, Digital Theology, Game Studies, Human Computer Interaction

1. Introduction

The intersection between videogames and religion was historically neglected (Campbell & Grieve, 2014, 2–3). However, the recently founded *International Academy for the Study of Gaming and Religion* (IASGAR), *Videogaming and Program Unit in the American Academy of Religion* (AAR), and the online journal, *Gamevironments* (all between 2014–2015), attest to the widespread import of digital games not only for entertainment, but also meaning-making for youth and adults across the globe and varied religious affiliations. In her chapter on "The Importance of Playing in Earnest," Rachel Wagner highlights some of the surprising symmetry between gaming and religious life, including but not limited to order-making mechanisms, predictive

Erin Raffety: Doing Theology *with* Videogames – Insights for Computational Theology, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 51–67. DOI: https://doi.org/10.11588/heibooks.1521.c21939

capacities, and escapist elements (2014, 193). But are videogames and religion merely symmetrical or metaphorical?

Even as the field of Digital Religion has taken up videogaming and even religious education and formation, there are few studies that look to videogames themselves or those who play them as a source of theological knowledge. In other words, theologians participate and consume, but we are not always critical theologically about the difference those actions make, nor do we frequently create our own models for further study. In order to assess the study of videogames for theological knowledge-making, let alone Computational Theology, this chapter begins by discussing and defining the various intersections between the fields of Digital Humanities, Computational Humanities, Digital Theology, and Computational Theology. I argue that Computational Theology, which creates and studies computational models, like algorithms and machine learning in videogaming, offers a significant opportunity for theological knowledge-making, so long as it does not become divorced from the users who make and utilize such models or obscure the researchers and research methods for studying them. As a practical theologian who uses qualitative methods, I use mixed-methods technological approaches, including, for instance, using digital tools such as Zoom, to gather qualitative data, and computational models, such as videogames, to study interactions between users and God. By drawing on background from *The Spiritual Loop Project,* a study of disabled videogamers and their Christian communities funded through the *Templeton World Charity Foundation* from 2020–2022, the article also elaborates three noteworthy lessons from our work in Computational Theology that may further conversation for the burgeoning field.[1]

2. Defining Computational Theology (and its Methods)

In his paper, "Ain't No Way Around It: Why We Need To Be Clear About What We Mean By 'Digital Humanities,'' Michael Piotrowski argues that new disciplines are not distinguished by new methods, but rather by "a particular combination of (1) a research object and (2) a research objective" (Piotrowski 2020, 10). Piotrowski continues, "…how does research – in whatever field – come to new insights? …the answer is easier than it may seem: it does so by *building models* of its research object" (Piotrowski 2020, 10). Piotrowski goes onto argue that as "universal modeling machines" (Piotrowski 2020, 11) computers are particularly adept at answering research questions. He concludes, "Thus, the difference between computational humanities and most traditional research in the humanities is not that computational humanities

1 This chapter focuses on distilling insights relevant to the field of computational theology from *The Spiritual Loop.* For a more thorough overview of *The Spiritual Loop* research methods, game design, and study results, see Raffety & Insa-Iglesias 2023.

constructs models, but that computational humanities construct models that can be manipulated by the computer, i.e. *computational models*, or more generally, *formal models*" (Piotrowski 2020, 11). As I understand it, Piotrowski wants to move computational research in the humanities toward a definition that actually meets three criteria, namely research of a (1) computational research object; (2) with a research objective related to computational humanities; and (3) facilitated by a computational model. This not only helps move the definition of Digital Humanities away from a simple rebranding of the traditional humanities, but also offers clarity as to how Computational Humanities research differs from humanities research by creating new knowledge using computers to produce models.

Before I move onto examine how we used *The Spiritual Loop* prototype as a computational model for Computational Theology, it is important to try to bring together Piotrowski's definition of the Computational Humanities with Peter Phillips, Kyle Schiefelbein-Guerrero, and Jonas Kurlberg's paper on "Defining Digital Theology" (Phillips et al. 2019). In their article, Phillips, Schiefelbein-Guerrero and Kurlberg outline some of the history of the development of Digital Humanities and the CODEC Research Centre at Durham's advocacy for a parallel "big tent" Digital Theology that encompasses a wide variety of projects and approaches and reflects "the prevalence and omnipresence of digitality" (Phillips et al. 2019, 33). Although theologians are as much impacted and conversant in digital technology as other scholars, Phillips, Schiefelbein-Guerrero and Kurlberg note confusion around the concept of Digital Theology, in contradistinction to the much more developed field of Digital Religion (Phillips et al. 2019, 33).

Drawing on established scholarship to map the waves of research in Digital Religion and the Digital Humanities, the authors conclude that these waves are both chronological and methodological (Phillips et al. 2019, 34–36), a distinction they further when developing their definition of Digital Religion as *purposively* "sociological and descriptive whereas the purpose of Digital Theology is theological" (Phillips et al. 2019, 37). They do note some crossover here, acknowledging that Digital Religion can be theological and presumably that Digital Theology can also be sociological (37), but ultimately (even though they do not specify it as I have below), they try to separate the two disciplines, much the way Piotrowski does, by way of research object: God for Digital Theology and religion for Digital Religion; and research objective: theological knowledge for Digital Theology and sociological knowledge (about religion) for Digital Religion.

Yet, this is where Piotrowski's third point, that disciplines develop new insights by building models of their research objects, in my opinion, demands some attention to methods. First, in looking back at Piotrowski's definition of Computational Humanities, we can see how wedded it is to the discipline of computer science, whose methods of coding and creating computing models with codes, are already inherent to the discipline. For instance, quoting from an inaugural use of the term, Computational Humanities (CH), Piotrowski writes, "[…] CH is a discipline that should provide an algorithmic foundation as a bridge between computer science and the humanities" (Biemann et al. 2014, 81, cited in Piotrowski 2020, 8). Therefore, embedded in the definition of Computational Humanities is this turn to computer science, whose methods, making computational models, although not new, are still integral to making such knowledge. Although Phillips, Schiefelbein-Guerrero and Kurlberg assert, "the disciplines of history and theology are located in the humanities, whereas anthropology and ritual studies are located in the social sciences" in their effort to explain the rough differences between digital theology and digital religion (Phillips et al. 2019, 37), notably, they make no mention of methods. However, in my mind this becomes a critical nexus in furthering understanding of what we are talking about when we are talking about Digital Theology.

Second and to explain further, I do not think it is fair to say theology, along with history, remains "located in the humanities, whereas anthropology and ritual studies are located in the social sciences," especially because contemporary studies of digital culture need consider and study just how human beings interact with technology to understand and make knowledge about God. Importantly, Phillips, Schiefelbein-Guerrero and Kurlberg acknowledge that theology is not just the study of God, God's interaction with the world, or the mystery of faith, but also includes, "thinking through that connection with the other" (Phillips et al. 2019, 37). Hence, it is critical to note that although many theologians studying God and seeking to make theological knowledge do so by studying ancient texts, history, and even evaluating and creating theological systems, other theologians study God and create theological knowledge by studying and evaluating human interaction with the divine. For instance, practical and pastoral theologians in the US, Europe, and elsewhere have argued persuasively that ethnographic, qualitative methods, and even quantitative methods can be approaches that help us to study how human beings create, interact with, and shape theological understandings in the world.[2]

Third, this is important because it suggests that Piotrowski's approach to computational humanities, one that places the emphasis on a particular method of model-building, will be too narrow for theology, a multi-methods discipline that is not only interested in modeling God, but also in understanding the interaction between God, users, and computational models. It is also important, because given the ubiquity of digital culture, digital theologians may necessarily use digital tools, such as Zoom, and computational models, such as videogames, as we did, for instance, to simultaneously study interactions between users and God. However, the current waves in digital theological research that Phillips, Schiefelbein-Guerrero and Kurlberg suggest (2019, 37–40) do not distinguish between these methods. Yet this is critical, not only because it is ethically vital for researchers to specify just what it is they are doing and how they are doing it, but also because good knowledge-making demands

2 See e.g. the work of *The Ethnography & Ecclesiology Network*, *The International Society for Empirical Research in Theology* [ISERT] and Ward & Tveitereid (2022).

epistemological clarity and methodological precision. As we continue to make theological knowledge with these mixed methods, both digital tools and computational models, we need to be reflexive about how the methods and models impact our knowledge-making.

In short, I do think that Computational Theology offers a significant point of development for theological knowledge, so long as it does not become divorced from the users who make and utilize computational models, nor the digital methods its researchers may use to study it. In other words, any definition of Computational Theology needs to enumerate how humans and methods interact with both God and technology. Failure to do so is unethical, imprecise, and even anti-theological, in that it obscures the ways by which we as humans come to know and understand God. I like that a focus on computational models moves us a bit afield from the "prophetic re-appraisal" Phillips et al. associate with progressive digital theology (Phillips et al. 2019, 39–40). Instead, by focusing on how researchers and people of faith are not just consumers but creators of computational models, we begin to ask good questions about not just what computers do but what we (and God) ultimately do with them. In the sections that follow, I enumerate what kind of lessons this sort of creative computational, yet relational work can have for theologians, with specific reference to our work with *The Spiritual Loop Project.*

3. Insights from *The Spiritual Loop Project*

3.1 Lesson 1

> Computational theologians wishing to engage videogames must center users in prototype development and methods for study.

From Fall 2020-Spring 2022, the *Center of Theological Inquiry,* funded by a grant from the *Templeton World Charity Foundation, Diverse Intelligences Initiative* (TWCF Grant Number 0265), and in collaboration with Glasgow Caledonian University in Scotland, created a *Minecraft* videogame prototype called, *The Spiritual Loop Project.* This videogame prototype was designed and developed for fostering spiritual growth and connection based on digital ethnographic research with disabled, neurodivergent[3] persons and their Christian faith communities in the United States. Considering the lack of access disabled persons experience with respect to Christian communities in the U.S. (Carter 2007), alongside the disproportionate emphasis on educational and

3 Neurodivergence is a broad term that can include autistic persons and persons with ADHD, as well as persons with dyslexia, Tourette's, other emotional and behavioral conditions, as well as persons with mental health diagnoses.

therapeutic outcomes with respect to neurodivergent gamers (Spiel & Gerling 2021), our participatory fieldwork with neurodivergent players led us to emphasize the game's opportunities for spiritual connection and growth versus mastery of biblical content or Christian virtues.[4]

Our guiding research question was, "Can machine learning be used to enhance the spiritual lives of disabled persons?"[5] Although this question sounds quite innocuous and straightforward, numerous disability scholars and activists have noted concerning prejudice in the way much artificial intelligence tends to assume an able-bodied user, thus reflecting and reinforcing ableist biases that threaten to further pathologize disability (Alper 2017; Nakamura 2019; Whittaker et al. 2019). The broad cultural bias expressed toward disability can actually be unknowingly and problematically integrated into machine intelligence in a way that perpetuates power asymmetries and further marginalizes disabled people (see e.g. anthropologist Karen Nakamura's work on the problem of self-driving cars running over people in wheelchairs, 2019). Both the pervasive nature of disability bias, and the intersectional and diverse nature of disability experiences make them particularly difficult to codify with respect to AI. A multi-author report on "Disability, Bias, and AI" identifies implicit biases in technology that tend to reinforce ableist ideas of normal and treat impairment as an object of repair or mediation, thus reinforcing a crude medical model of disability (Whittaker et al. 2019).

Although not much research has been done on virtual worlds and multiplayer videogames from disabled perspectives, the advent of game streaming with platforms such as Facebook and Twitch has substantially widened the social component of gaming. For instance, Kathryn Ringland's extensive ethnographic work with the "Autcraft" community, a community of autistic players in the online game *Minecraft,* demonstrates the variety of forms of technology and platforms that foster simultaneous connection and communication, as well as the importance of this virtual world as social space for autistic youth (Ringland 2019a). Ringland's work, in fact, highlights how important it is that theological researchers aiming to do digitally reflexive computational work do not proceed in a vacuum, but make use of other epistemologically analogous fields. For instance, the field of *Human Computer Interaction,* or HCI in the United States, not only centers interactions between users and computational models,[6] but offers a critical perspective on ethics and personhood, through its robust

4 Portions of this chapter also appear in Raffety & Insa-Iglesias (2023).

5 The article in footnote 4 further explores this intersection between machine learning, spirituality, and neurodivergence.

6 In their book *Digital Theology: A Computer Science Perspective* (2021), Erkki Sutinen and Anthony-Paul Cooper argue for an interactive design process for developing technology for faith communities, that comes out of computer science and "follow the principles of co-design…[in which] diverse stakeholders work together throughout the design process" (Chp 1, Section 1.2). Although there seems to be much resonance between this approach and HCI, especially because HCI has

consideration of disabled users (see e. g. Ringland et al. 2016; Ringland 2019b; Williams et al. 2021; Williams & Gilbert 2019). Not only has work in this area been critical for my own epistemological and methodological learning in research with disabled users, but it is scholarship that theologians, given their applied focus, cannot afford to overlook. It demonstrates how to use digital methods to study computational models, and while it does not primarily seek to produce knowledge about God, of course, the discipline has already thoughtfully mapped relationships between researchers, users, and machines.

Hence, our research design centered disabled users, particularly neurodivergent persons, as experts, relying on fieldwork with them and their communities to shape the development of a videogame prototype to test the insights for machine intelligence in enhancing spiritual lives.

For instance, in Winter 2020, I began fieldwork by observing the majority of the neurodivergent gamers on Zoom playing their favorite games. This helped me experience the features disabled gamers particularly enjoyed so that I could work to incorporate them into the future prototype and offered a familiar medium with which to do so: gamers could talk over Zoom and tell me what they were doing, but they could also simply allow me to observe, type comments in the chat, or offer verbal exclamations as they played. I also conducted focus group sessions via Zoom that Winter to provide insight on what disabled gamers and their respective Christian communities wanted to see in a game prototype. This pre-design fieldwork, conducted through the digital platform of Zoom, gathered critical insights about what these communities and their disabled congregants valued in both videogaming and church life.

In Spring 2021, the technology fellow worked to construct a novel game in *Minecraft* that met the specifications developed from these pre-design fieldwork and focus group sessions. Here we should note that our process fell short of the principles of co-design (Sutinen & Cooper 2021) in that, given constraints of time and expertise, our technology fellow was responsible for implementing the design of the game. Yet, centering disabled users in the study also required that we worked with them to create methods that allowed us to access their experiences playing the game. In Summer 2021, the research units had an opportunity to test various elements of the game and provide feedback. Although I tested other gaming platforms such as Discord, due to research participants' widespread familiarity with Zoom and preference for visual and audio communication during play, the technology fellow and I developed a method for online gameplay that involved simultaneous Zooming for communication and data collection during play. We also tested this approach in Summer 2021 orientation sessions, during which participants were instructed over Zoom how to download the current version of *Minecraft*, log onto the server, and periodically share their screen to demonstrate challenges or observe other players' play. These orientation sessions

thoughtfully centered disabled users in design and studied those users' experience, I cite that literature here.

also allowed research participants to give some initial feedback on some of the gaming elements, as the technology fellow was still working to complete the prototype through August 2021.

From Sept 2021–February 2022, each research unit played the game in their units, along with me. Finally, in January and February 2022, each research unit participated in a feedback session with me, where they provided verbal and chat feedback on their experience playing the game. It should be noted that research units played exclusively with me and the other members of their unit so that I could observe how the game impacted spiritual play, conversations, and relationships among persons who already had prior relationships.

What I believe this rather exhaustive description of the methods in the project demonstrates is that methods for studying computational models must also center users and work in tandem with the models themselves. First, as I have shown, without centering users in research design and development, it is far too easy for existing biases to be imported into computational models, particularly models utilizing artificial intelligence. However, because theology is studying not just God but the interaction between God and human beings, mixed methods approaches, like the one I present above, that use both computational models and digital technology to study how users interact with these models, need more development and transparency in the digital theology literature in order to substantiate epistemological and theological insights. Finally, computational theology ought to both critique and refine relevant fields, such as HCI, as it continues to refine its methods toward developing both computational models and methods for evaluating their import.

3.2 Lesson 2

> Computational theologians must appreciate play as a site of theological knowledge-making, moving from observing structural symmetry between religion and games to doing theology with gamers themselves.

This brings me to my second point, that for computational models to make theological contributions, we must seek to truly integrate theology in research, rather than just observe or identify structural symmetry between religion and videogames. Much of the existing work on videogames and religion tends to fall into two camps: (1) sociological work that observes the surprising symmetry between gaming and religion (Campbell & Grieve 2014; Wagner 2014); or (2) studies of how didactic games make educational contributions to particular religious communities (Gottlieb 2015; Hutchings 2023). Although a few recent publications helpfully complicate these categories (Garner 2021; Hess 2019), the existing binary demonstrates a tendency to focus on the computational models themselves, rather than interaction between user and model. Indeed, these perspectives tend to epistemologically undermine the agency of

Fig. 1 This figure shows a bird's eye view of *The Spiritual Loop Project Minecraft* server. The first environment where participants start playing is the village, which includes the fountain square, plots (yellow house plot and mural plot), and the church. This server is inhabited by Non-Player Characters who guide players, promote social interaction, and collaborative play to win the game.

gameplayers, rather than insist on their agency in not just consuming the game, but creating new forms, in our case, of theological knowledge. Both for Christians and for disabled persons, play is often instrumentalized for religious or educational purposes, thus ironically subordinating or controlling play itself (see Raffety & Insa-Iglesias 2023; Spiel & Gerling 2021). Therefore, in centering disabled users and harnessing *Minecraft* "maker culture" (Ringland 2017), we shift the theological emphasis from the model, or the game itself, to how users, through "play" with theology, offer critical insights for Christian communities.

Owing to our players' interest in creating a game that allowed them to simulate worship, we created a village style game with a Christian storyline. The game takes place in a small village (see Fig. 1) consisting of a main square, with a fountain, several house plots, and a small church with a bell tower. The game's purpose is to cooperate with players to complete a set of tasks (individual and cooperative), called "advancements" in *Minecraft* on each level, and make it to the last level to "win the game" (see Fig. 2 (B) for level 0 advancements and (C) for level 1 advancements). The advancements are designed to encourage interaction and cooperation and correspond to Christian biblical themes and principles. When players complete all the advancements, they are invited to participate in the great feast, a banquet that simulates the last supper. Upon completion of this final level, they advance to creative mode, where they are given access to all resources and can explore beyond the pre-existing village, simulating heavenly freedom.

Fig. 2 This figure shows some scenes from the *Minecraft* server. Players start the game next to the fountain square (A), where they find a book with instructions about how to play, suggested by the chat. Players need to complete a set of individual and collaborative tasks (called "advancements" in Minecraft) that are listed in the book or can be visualized on the advancement tab: level 0 (B) and level 1 (C). The NPC, AI witness, guides players through the game and provide hints when interacting with it, for example, on the mural plot (D) or at the Community House (E).

The *individualized tasks* are custom advancements where each player is required to interact with elements of the game or perform tasks that benefit the community. For example, the task "find your chest" requires players to find a chest labeled with their name; the task "build your house" requires players to place a minimum number of blocks into the configuration of a house on their plot of land to welcome others and interact with them; the task "speak to your neighbors" requires socializing and interacting with others in the game through utilizing the chat feature. Other individualized tasks include "discover the mural," "call to worship," "visit the church," and "light the church." The *cooperative tasks* are customized advancements where cooperative play is required to benefit the community. For example, the task "share to care" requires sharing resources with others; the task "cooperate to discover the mural" requires cooperating to break blocks to discover the village mural. Players are not able to advance through the game if they do not discover the meaningful cooperation necessary to complete the tasks (see the hint provided by NPC in Fig. 2 (D) and (E)). Other cooperative tasks include "share time together," "worship together," and "the great feast."

Although a few of the advancements can be individually completed (i.e. "find your chest," "build your house," etc.), most advancements require cooperative action to be completed (i.e. "share to care," "worship together," etc.). In fact, many cooperative advancements, when not completed together, have mechanisms built into the game to slow individual actions and prevent players from advancing, simulating the Christian doctrine of sin. For instance, we used white witness characters (non-player characters or NPCs) designed with decision trees who appear strategically in the game when players were taking nonproductive actions to provide hints, biblical/spiritual advice, or assistance regarding how to process through the advancements. The implementation of the witnesses follows decision trees where the NPC decides based on a set of conditions.

There were several findings with respect to gameplay that are generally relevant to demonstrating how studying computational models and users can facilitate theological knowledge-making. First, despite the manufactured cooperative elements in the game, players expressed an even greater desire to collaborate with other players. For instance, many players named cooperative tasks built into the advancements, including discovering the mural, worshiping in the church, sharing resources, and eating together, as highlights of the game experience from both a social and spiritual perspective. However, many players also explicitly expressed a desire to collaborate on individual advancements, such as "build your house," which were not supported by the existing game. From a Christian theological perspective, this suggests that play itself can expand our notions of what Christian fellowship, care, and community-building can even look like. We assumed that each player would want autonomy over their own house, but the feedback suggesting that element be cooperative, attests to more challenging Christian teachings, regarding having all things in common (Acts 2:44–45), for instance, that are often either inaccessible within our individualist, capitalist culture, or simply ignored. This simple insight shows the way in which the interaction between the user and the computational model, the videogame, develops theological claims and practices that are relevant to contemporary Christian worship, culture, and community.

Another surprising finding was that even the players who had ample knowledge of *Minecraft* and led with expertise, noted that more instruction in the game would still benefit both them and other players. In this excerpt, Player A, an inexperienced neurotypical player and Player B, an experienced neurodivergent player, agree that there wasn't enough direction in the game:

> Player A (NT Player): For me, there wasn't enough direction. I didn't really know like what I was doing…and when it ended, I was surprised because I felt like I didn't really know what I had done. Like, I couldn't remember a few tasks, but there were other things that just happened. Or I was told to press this button and a little banner popped up and then we at the end, uh, had made it, uh, as like an inexperienced player like that, I was missing some of that.
>
> Player B (ND Player): [There was] too little direction [in the game]. There should've been a little more like push for them. Some things for the um, what was it? The…dinner thing. It should have been more straightforward on like where we had to be at what time. (Feedback Session, January 27, 2022)

The surprising desire across players with different levels of gaming experience for more direction in the game suggests that the merit of play is not necessarily amplified by challenge and struggle and can coexist with meaningful support, clarity, and assistance. Of course, it is impossible to say whether this insight is specific to Christian gaming or Christian gaming communities, but its suggestion that accessibility may be an integral and often overlooked aspect of play and connection, dovetails so strikingly with the American culture of competitiveness in which this game was played.

Finally, as nearly all the players also suggested that the game did not need to have a logical end point, but rather should have gone on and on, we realize that these Christian gamers are keenly focused on cooperation, accessibility, and are much more comfortable with ambiguity than we might have expected. Those often overlooked aspects of play are not merely symmetrical to the model itself, but insights that are uniquely and specifically drawn out of agentive play with the model. These are just a few examples of how computational models may produce new theological insights, demonstrating not just the importance of models themselves but the perceptions and insights that gamers enact with them in the scope of gameplay that may have theological implications even beyond the game.

3.3 Lesson 3

> Computational theologians must not mistake games for neutral objects; rather, they must interrogate the theological underpinnings of computational models.

Another critical insight from our study was that despite centering users in our prototype design and study methods, it was only in the feedback loop from users that we gleaned some of the theological implications of the models themselves. In our case, we chose to build our existing game within the *Minecraft* videogame because of its widespread popularity, accessibility, and adaptability. However, by building within an existing game, we inherited many of the game's features, storyline, and culture. As aforementioned, part of this culture is a "maker" culture, where players can design and build structures, and alter the existing environment. Yet, the game is also organized around individual advancements where players gradually acquire resources to build tools to access and alter their environment. Players noted that not only did our cooperative tasks rather conflict with some of the individual elements of the game, but our cooperative tasks also introduced competing values into the existing game. As one player remarked quizzically, "[We're] building an experience for, spirituality and collaboration, and the first thing it tells you is how to make swords" (RU1 February 22, 2022). Here the player implies that she experienced some tension between a game that wants to foster Christian spirituality and collaboration, yet contains existing programming to prompt and encourage players to create weapons to defend themselves and potentially harm others.

Beyond mere symmetry between religious and videogame themes, is it reasonable to assume that videogames themselves may harbor their own theologies? In her book, *Christian Ethics for a Digital Society* (Ott 2019), Kate Ott makes this point with regards to the ideologies of social media platforms, but through our fieldwork, users highlighted the theological implications of *Minecraft*'s features, storyline, and culture. As the player above went onto say, "You're defending your crops and it's like, no in our game, we would, what is it? We would mold our swords into plowshares. And we would give the crops away…we would bend it to our narrative" (RU1 February 22, 2022). In her comments, this player points out the extent to which our computational model offered a parallel experience, rather than an integrated experience, in *Minecraft.* Although we offered an alternate storyline in *The Spiritual Loop,* it was not integrated into the existing culture of *Minecraft,* such that it even caused theological tension for players. Yet, the player sees potential within the existing game's infrastructure to offer theological meaning, by creating, for instance, cooperative advancements that offer players opportunities to beat those swords into plowshares in accordance with the ancient scriptures (Isaiah 2:4).

These players' insights clarify that computational theologians must learn iteratively, toggling between games and gamers, in order to fully understand the theological implications of these interactions. As one research participant put it,

> …Christianity by and large is in this country a very independent, non-collaborative endeavor. [So] maybe instead of trying to teach that endeavor through *Minecraft,* we should be teaching Christianity how to be collaborative…through *Minecraft.* Maybe we should be learning from *Minecraft* rather than using it as a tool to convey something that's already not quite biblical, but this is just the way we accept things are. (RU3 Feedback Session, Jan. 18, 2022)

Although a bit abstract, the player in this quote seems to be suggesting that Christianity in the United States has appropriated values of independence from the culture that run counter to the Gospel. Hence, if we create computational models with this version of Christianity in mind, we necessarily fall short, encounter, or even merely recreate the same biases. However, the player suggests that by learning from the collaborative modes and features inherent to videogaming, or in this case in *Minecraft,* perhaps we can even open up more faithful practices of Christianity. This player's insight is important because it brings together the three lessons in this chapter: first, it highlights how important it is for computational theologians to move beyond observing symmetry between gaming and Christian practices and study gaming's theological implications with seriousness and resolve. Second, it points out how this study on its own, though, is not enough. Rather, in seeking to learn from computational models, computational theologians must take users seriously as not only as mere players but as theologians in their own rights, whose contributions and insights in gaming, can lead to valuable theological knowledge-making today.

4. Conclusion

This chapter has explored the potential for Computational Theology to distinguish between digital methods for studying computational models and computational models themselves, such as videogames, by clarifying how users interact with such models and what it teaches us about God. By drawing on findings from *The Spiritual Loop Project,* I show how centering users in the making and studying of videogames can allow us to think theologically with games and gamers themselves. This helps move theological knowledge-making beyond mere analysis of models toward playing with and learning from models, demonstrating tremendous potential for the field of computational theology when it comes to studying videogames.

References

Alper, M. (2017). *Giving Voice. Mobile Communication, Disability, and Inequality.* Massachusetts: MIT.

Campbell, H.A. & Grieve, G.P. (2014). *Playing with Religion in Digital Games.* Bloomington: Indiana University Press.

Garner, S. (2020). Sacred Pilgrimage in Playful, Digital Spaces. In J. Tucker & P. Halstead (Eds.), *Sports and Play in Christian Theology* (pp. 93–108). Lanham, MD: Lexington Books/Fortress Academic.

Gottlieb, O. (2015). Current Key Perspectives in Video Gaming and Religion. Theses by Owen Gottlieb, *Gamevironments,* 1(3), 18–25. URL: https://journals.suub.uni-bremen.de/index.php/gamevironments/article/view/90/76 (Accessed: 21 June 2024).

Grieve, G., Radde-Antweiler, K., & Zeiler, X. (Eds.). (2015). Special Issue. Current Key Perspectives in Video Gaming and Religion, *Gamevironments,* 1(3), 1–10. URL: https://journals.suub.uni-bremen.de/index.php/gamevironments/article/view/126/112 (Accessed: 21 June 2024).

Hess, M.E. (2019). Why Games and Gaming Might Be the Best Way and Place in Which to Consider the Meaning and Purposes of Theological Education. A Reflection, *CrossCurrents,* 69 (1), 80–94.

Hutchings, T., (Ed.). (2023). Teaching with Games. Formative Gaming in Religion, Philosophy and Ethics, *Gamevironments,* 19, 1–18. URL: https://journals.suub.uni-bremen.de/index.php/gamevironments/article/view/243/193 (Accessed: 21 June 2024).

Nakamura, K. (2019). My Algorithms Have Determined You’re Not Human. AI-ML, Reverse Turing-Tests, and the Disability Experience. In *Proceedings of the 21st International ACM SIGACCESS Conference* (1–2). Pittsburgh PA: Association for Computing Machinery.

Ott, K. (2019). *Christian Ethics for a Digital Society.* London: Rowan & Littlefield Publishers.

Phillips, P., Schiefelbein-Guerrero, K., & Kurlberg, J. (2019). Defining Digital Theology. Digital Humanities, Digital Religion, and the Particular Work of the CODEC Research Centre and Network, *Open Theology,* 5, 29–43. DOI: https://doi.org/10.1515/opth-2019-0003 (Accessed: 21 June 2024).

Piotrowski, M. (2020). Ain’t No Way Around It. Why We Need to Be Clear about What We Mean by “Digital Humanities”. Online: *SocArXiv* (pp. 1–16). DOI: https://doi.org/10.31235/osf.io/d2kb6 (Accessed: 21 June 2024).

Raffety, E. & Insa-Iglesias, M. (2023). Re-imagining Christian Education through Neurodivergent Fellowship, Play, and Leadership in Online Videogaming, *Gamevironments,* 19, 80–114. URL: https://journals.suub.uni-bremen.de/index.php/gamevironments/article/view/220/187 (Accessed: 21 June 2024).

Ringland, K.E. (2017). Who has Access? Making Accessible Play Spaces in Minecraft for Children with Autism, *Analog Game Studies,* 4(3). URL: http://analoggamestudies.org/2017/05/who-has-access-making-accessible-play-spaces-in-minecraft-for-children-with-autism (Accessed: 21 June 2024).
Ead. (2019a). 'Autsome'. Fostering an Autistic Identity in an Online Minecraft Community for Youth with Autism. In N. Greene Taylor, C. Christian-Lamb, M.H. Martin & B. Nardi (Eds.), *Information in Contemporary Society. 14th International Conference, iConference 2019. Proceedings* (pp. 132–143). Washington, DC: Springer [= *Lecture Notes in Computer Science,* 11420].
Ead. (2019b). A Place to Play. The (Dis)Abled Embodied Experience for Autistic Children in Online Play Spaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Glasgow: Association for Computing Machinery.
Ead., Wolf, Ch.T., Boyd, L.E., Baldwin, M.S. & Hayes, G.R. (2016). Would you be mine. Appropriating minecraft as an assistive technology for youth with autism. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility 2016* (pp. 33–41). Reno Nevada: Association for Computing Machinery.
Spiel, K. & Gerling, K. (2020). The Purpose of Play. How HCI Research Fails Neurodivergent Populations, *Association for Computing Machinery (ACM),* 1(1), 1–39.
Sutinen, E., & Cooper, A.-P. (2021). *Digital Theology. A Computer Science Perspective.* Bingley: Emerald Publishing Limited.
Wagner, R. (2014). The Importance of Playing in Earnest. In H.A. Campbell & G.P. Grieve (Eds.), *Playing with Religion in Digital Games* (pp. 192–213). Bloomington: Indiana University Press.
Ward, P., & Tveitereid, K. (Eds.). (2022). *The Wiley Blackwell Companion to Theology and Qualitative Research.* Hoboken: Wiley Blackwell.
Whittaker, M., Alper, M., Bennett, C.L., Hendren, S., Kaziunas, E., Mills, M., Ringel Morris, M., Rankin, J.L., Rogers, E., Salas, M., & Myers West, S. (2019). Disability, Bias & AI Report. New York: *AI Now Institute.* URL: https://ainowinstitute.org/publication/disabilitybiasai-2019 (Accessed: 21 June 2024).
Williams, R.M., & Gilbert, J.E. (2019). 'Nothing About Us Without Us' Transforming Participatory Research and Ethics in Human Systems Engineering. In R.D. Roscoe, E.K. Chiou & A.R. Wooldridge (Eds.), *Advancing diversity, inclusion, and social justice through human systems engineering* (pp. 113–134). London: CRC Press.
Williams, R.M., Ringland, K., Gibson, A., Mandala, M., Maibaum, A., & Guerreiro, T. (2021). Articulations Toward a Crip HCI, *Interactions Dialogues (Conversations),* 29–30.

Figure Credits

All figures show screenshots of the *Minecraft* servers from *The Spiritual Loop Project.*

Fig. 1: The screenshot was taken by the author herself.

Fig. 2: The screenshot was taken by Maria Insa Iglesia, Technology Fellow in *The Spiritual Loop Project.*

On the Practice of Computational Theology

Christopher A. Nunn

https://orcid.org/0000-0001-7208-8636

Abstract This essay provides an overview of the recent research field "Computational Theology." The terminology associated with this field will first be examined in more detail. The task of Computational Theology is to address theological research questions using the tools found in the Digital Humanities. However, in this context, what does Digital Humanities mean and what concept of theology funds this complex of phenomena? Computational Theology and Digital Theology come from different academic fields. Consideration of concrete practices shapes the research field itself. Thus, in the second part of the essay, a field analysis is performed to glean insight into the concept of Computational Theology.*

Keywords Computational Theology, Digital Theology, Theology, Religious Studies, Digital Humanities, Computational Humanities

1. Doing Computational Theology? A Survey of the Field

1.1 Terminology

Belonging to the field of Computational Theology is primarily a claim about methodology and a description of certain research practices. Theological questions are dealt with by means of Digital Humanities (DH). The ambivalent concept of DH can be specified as aligning more with what Ramsay labels type 1 as opposed to type 2:

> Ramsay argues that Type 1 digital humanities (DH) is 'united not by objects of study, per se, but by a set of practices that most regarded as intimately related: text encoding, archive creation, text analysis, historical gis, 3d modeling of archaeological sites, art historical cataloging, visualization,' and general meditation on what all of these new 'affordances might mean for the study of the human record.' He explained that 'Type 1 DH is [a] community' and 'in early 2001, this community fatefully decided to call itself

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

Christopher A. Nunn: On the Practice of Computational Theology, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 69–95. DOI: https://doi.org/10.11588/heibooks.1521.c21940

> "digital humanities,"' as humanities computing sounded like a 'campus technical support group.' Ramsay argues that 'digital humanities' is 'useful because it distinguished our activity from media.'
>
> With Type 2, on the other hand, Ramsay argues that 'I don't know exactly how it happened […] Media studies practitioners were digital humanists; people who had devoted several decades to digital pedagogy were digital humanists; cultural critics who were interested in Internet culture were digital humanists; and digital artists of a certain variety were digital humanists.' The resultant confusion of disciplinary identity, for Ramsay 'sounds like the recreation of the humanities itself after some technological event horizon.' Type 2 digital humanities, then, is a more expansive notion, including media theory, cultural critique, media and communications, etc. (Berry & Fagerjord 2017, 36f., which refers to a no longer existent Blog post by Stephen Ramsay).

Computational Theology thus refers to a sub-area of Digital Theology,[1] which does not necessarily have to be classified as Computational Theology, as Digital Theology can also be used to practice type 2 of the DH. Computational Theology can also overlap with the field of Digital Religion, such as when religion in social media is examined with the help of programming techniques.[2] In practice, however, the goals of Digital Religion are much broader. The term Digital Religion is used to describe and reflect on "religious practices in the digital realm" (van Oorschot 2023, 17).[3] The focus of this broader field is much more on contemporary theology. Raffety's perspective in the previous chapter can also be situated in the field of Digital Religion (p. 55): "[A]ny definition of computational theology needs to enumerate how humans and methods interact with both God and technology. Failure to do so is unethical, imprecise, and even anti-theological." According to Raffety, Computational Theology is a subcategory of theology in the digital realm. However, this interpretation limits the potential epistemological scope of this research approach, as Karcher (2020, 133) asserts in a definition of Digital Theology:

The problem with this approach is […] the narrow understanding of theology, which is focused exclusively on religious phenomena in the digital space and the religious practice. If digital theology is defined exclusively as a new way of reflecting exclusively on practical theological phenomena as theology of/in the digital, all other theological disciplines – if not excluded *per se* from digital work – are forced to adopt a religious-practical dimension.

1 That Computational Theology occurs within the context of Digital Theology is particularly evident in the methods presented in Sutinen & Cooper (2021, 61–90), which can also be largely attributed to the repertoire of the DH methods.

2 On a machine-learning approach to analysing the use of religious Twitter hashtags, see Veidlinger ([2]2022, 132–140).

3 Cf. Campbell & Tsuria ([2]2022) for a comprehensive overview of Digital Religion.

The different understandings of theology presented by Raffety and Karcher is a result of diverging academic backgrounds. Karcher's European perspective on theology is influenced by Schnelle (2021, 39): "Theology is the academic (orig. "wissenschaftliche") reflection on the content and practice of a religion, the consideration and thought about religious claims about God, humanity, and the world." Religion is therefore primarily the object of study of an academic discipline. From an Anglo-Saxon perspective, the objective of theology could be described as "religious studies," which is distinguished from "theology:"

> [T]heology is a study – of something else, say, God, or of how to talk about God, or of how God talks. And it is the study of God in the Latin sense, with passion – for to 'study' theology in the primary sense of the expression is to do theology […] you don't practise religion by studying it, as you practise Islam by doing Islamic theology, or practise Christianity by doing Christian theology (Turner 2005, 26).

A religiously practical dimension is inherent to Anglo-Saxon Digital Theology.[4] Understanding the inherent practical aspects makes Raffety's claim understandable. From within this tradition, Rafetty also applies a similar standard to Computational Theology. Consequently, Piotrowski's understanding of the relationship between DH and Computational Humanities presented in this volume can also be expressed in the internal theological discourse. The term "Computational Theology" is not only based on the Computational Humanities in that the technical aspects are emphasized, but the differences result analogously from different academic cultures. At this point, however, two potential objections must be addressed that could call into question the meaning of new terminology.

1. Why not use "Digital Religious Studies" instead of Computational Theology? In the German-speaking scholarship, this would likely result in further terminological confusion: Religious Studies refers to *Religionswissenschaft,* which as a sub-discipline within theology, is geared towards theological inter-religious competence (cf. Feldtkeller 2006, 123), or, as a discipline completely independent of theology, empirically researches religions and religious phenomena (cf. Moenikes 1997, 197).

4 Cf. Sutinen & Cooper (2021, 13): "Digital Theology is, fundamentally, an academic subject deeply rooted in practical applications." The book *Digital Theology: A Computer Science Perspective* gives a concise overview of the research field of Digital Theology. The definitions by Phillips et al. (2019), which Karcher has in mind in his critique of them being too oriented to practical theology, are evaluated in a contrasting manner by Sutinen & Cooper (2021, 16): "[T]hey do not fit comfortably. This is largely because the Phillips et al. definition is focussed on theory and conceptual argumentation; it does not easily cover the more practical aspects of Digital Theology which have emerged over recent years." Different academic cultures can thus be seen from the critiques themselves.

2. Using the German term *"Digitale Theologie"* would convey the European understanding of academic theology as understood by Karcher. So why is adding yet another term necessary? Apart from the lack of global connectivity the use of German implies, the problem is that *Digitale Theologie,* introduced as a term in the German-speaking world in 2015 by Johanna Haberer, a theologian with a primary focus on media studies, has already been very strongly influenced by the English-speaking discourse (cf. van Oorschot 2023, 14). The following (roughly sketched) diagram (Fig. 1) illustrates the diversity of the different areas of Digital Theology, each of which stands on its own, despite some eventual overlap (van Oorschot 2023 provides a detailed guide). The establishment of a Computational Theology should therefore also be seen as an important reform effort to be able to name theological research in the sense of type 1 of DH and thus also be identifiable for those outside the field.

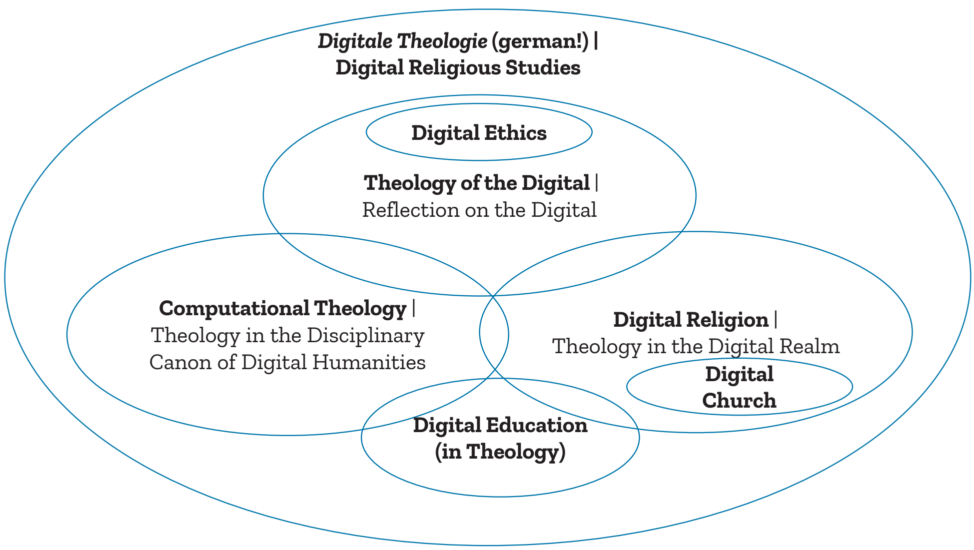


Fig. 1 Spherical Model of the different dimensions of Digital Theology

1.2 Cognitive Interests

Methodologically, Raffety makes a commendable contribution to Digital Theology. However, this version cannot be seen as Computational Theology, as it follows an understanding of DH according to type 2 and would presumably be categorized by Ramsay as "media and communications." Raffety's proposal is in line with the trend in Game Studies, where studies are focused on the "humanities of the digital," i.e., a reflection on the digital, as Burghardt (2024, 1) indicates. Burghardt also includes approaches to Game Studies that would correspond to a DH type 1 on p. 1f.[5] Video game technology is by no means excluded as a subject of Computational Theology and could be an additional chapter of the compendium in the future *living handbook.*

Scholars of Computational Theology can have varied approaches to the connection between theological questions and IT practices. For example, one can ask whether established theological methods (e.g., in the field of historical critical exegesis)[6] can be operationalized digitally. Conversely, established methods of the DH community could be examined to see whether they can be adapted for theological questions (the present volume could be an example of this). It would also be conceivable to develop new digital methods for theological questions, especially in an interdisciplinary work with scholars in the computer sciences.

2. Doing Computational Theology! Examples from Practice

After a lecture at the DHd2024 in Passau, during which I illustrated the lack of perception of theologians on the part of the DH community using Sahle's sphere model on the disciplines of DH (for this model, see p. 15 in the introduction to this volume), he sent me a modified version (Fig. 2), which now also includes theology. In fact, he already wrote in 2017 (11, n. 4):

> The diagram is very oversimplified. The presented "subject areas" are not all-encompassing but are simply representative. Other disciplines could be added and positioned in a certain place, understood as sub-disciplines of individual disciplines, or even shown to be an intersection between major disciplines. In this respect, some might also locate biblical studies differently, as part of theology (which is not included here but could perhaps fall under the similar category as philosophy), as an area that integrates methods from different disciplines, or as a separate discipline. Positioning the subject areas also shows the variations in orientations to different objects: abstract concepts (philosophy), language (linguistics), texts (literary studies) or objects in their history (history), and visual (art history) or material (archeology) dimensions.

Sahle therefore also grants theological research a place in the DH's canon of research fields. However, the position of theology does not align with theology's high level of interdisciplinarity. A focus on abstract concepts is too narrow; theology is also similar

5 See also Ensslin (2021).

6 See e.g. Al-Suadi (2021, 66), who compares the methodological steps of historical critical exegesis with corresponding counterparts of digital historical critical exegesis.

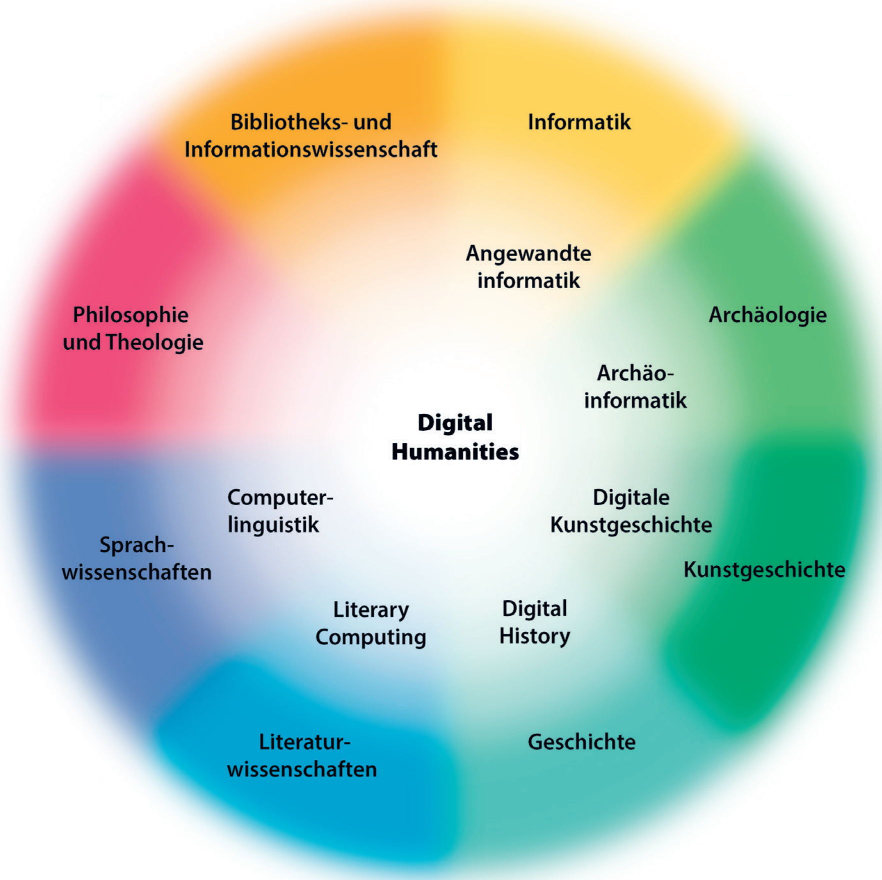


Fig. 2 Spheres of the Digital Humanities (modified)

to the other dimensions mentioned. Under this premise, however, there is no place within the diagram that would be appropriate. There would always be a deficiency somewhere. But apart from that, one might also ask whether it is appropriate for theology to be included in this context. To answer this question, we will now identify examples (!) of theological anthologies, special issues, and places of activity that can be assigned to Computational Theology (at least in part). While only a representative cross-section can be shown in this chapter, it is intended that the following information (along with the list of digital editions in the appendix of this article) will be supplemented or updated as part of the *living handbook.*

2.1 Anthologies and Special Issues Engaged with Computational Theology

Some emphasize (such as a few church historians) that DH research has been practiced in their own discipline for many decades. While the claim need not disputed here, one must keep in mind that these practices often involve research in neighboring disciplines on topics related to religion. Anderson (2019, 76) claims:

> Theologians have shown scant interest to this point in the tools for linking data, mapping, network analysis, text mining, and visualizing information that are fueling digital scholarship in other disciplines. My suspicion is that theological scholars may appreciate what their colleagues in other disciplines are doing but see them as irrelevant to theological inquiry.[7]

The citation from Clark & McBride Lindsey (2022, 16) must also be read in this light:

> Tim Hutchings observed that "religion is at best a marginal theme in digital humanities conferences and debates" (Hutchings 2015, 283). In recent years the Association of Digital Humanities Organizations – the largest DH scholarly community – has hosted panels on religion and theology and the American Academy of Religion now facilitates a number of DH-themed panels and papers at its annual meetings.

To name another example: the religious topics at the DH2023 in Graz were mostly presented by philosophers and historians – with the exceptions of the religious studies scholar Martin Prell and the new testament scholar and theological library specialist Juan Garcés at the panel: *Transforming the Pietist Tradition: Disciplinary Innovation through Linked Digital Engagement* and my presentation of this compendium – (cf. the *Book of Abstracts* of this ADHO Conference in Baillot et al. [2023]). In the following one should note that the contributions to the field from the various research are ensured to be theologians or religious studies scholars.

2.1.1 Computational Theology in Biblical Exegesis ("Digital Biblical Studies")

Most of the secondary literature even tangentially related to Computational Theology is likely found in biblical studies, especially in the area of the New Testament. Clivaz (2020, 98) provides a concise overview of the literature. One immediately notices that

7 Alternatively, theological research relies on the portals of neighboring disciplines with the risk that subject-specific potentials are not fully exploited, see Zahnd (2020, 115).

the contributions are mostly in the field of digital text criticism and editing.[8] In the special issue *Digital Humanities in Biblical Studies and Theology* (*Open Theology* 5(1) 2019), which was also co-edited by Clivaz, eight of the eleven contributions are devoted to this topic. Further approaches, such as those on redaction criticism, can be found in Al-Suadi & Ulshöfer (2021), on network analysis in particular in Czachesz (2022, 9–26), the subsequent six articles in a special issue of the journal *Annali di Storia dell'Esegesi* entitled *Network Science in Biblical Studies*, and on computer-assisted stylometric methods for authenticating the Pauline letters by Jannidis in this volume.

Garcés & Heilmann (2017, 29–52) provide an excellent introduction to the intersection of DH and biblical exegesis. In addition to methods of textual criticism and digital editing, they also devote a subchapter to stylometry, co-occurrence analysis, and digital intertextuality in New Testament research.

Biblical archaeology is a particular case that connects biblical studies with the DH. Collinet (2021, 153–156) outlines the potentials and problems of digital methods in this field.

2.1.2 Computational Theology in Church History

From the perspective of church history, the practices of Computational Theology are particularly evident in the creation and use of full-text databases and digital editions (cf. the appendix by Annette von Stockhausen at the end of this chapter).[9] This is due to the fact that their practice is only possible if suitable objects of investigation are available (cf. Heilmann 2022, 13 f.). The digitization, processing, and preservation of theologically relevant sources and artifacts thus always remains one (!) of the central concerns of Computational Theology. In 2020, von Stockhausen published a special issue made up of four articles in the *Journal of Ancient Christianity* on the patristics (and the New Testament), in which various (differently understood) editing projects are presented on the one hand, and techniques, theories, and guidelines relating to editing are discussed on the other:

- M. Cassin: Où en est l'édition de textes patristiques grecs aujourd'hui? Théories, méthodes et pratiques (11–59)

8 Fischer (2019, 203–219) illustrates the enormous advantages a corresponding prepared text critical apparatus brings to a digital edition.

9 Digital prosopographies should also be mentioned, yet there were no (or at least hardly any) theologians involved in the development (at least in the digital stage). On the developmental history of the *Prosopography of the Byzantine World* (PBW) and *Prosopography of Anglo Saxon England* (PASE), see Ch. Roueché et al. (2023).

- H. Strutwolf: Die *Editio Critica Maior* des griechischen Neuen Testaments – Editionsprinzipien, Editionstechnik und Digital Humanities (60–108)
- J. Walters: The *Digital Syriac* Corpus: A Digital Repository for Syriac Texts (109–122)
- A. von Stockhausen: Die Modellierung kritischer Editionen im digitalen Zeitalter (123–160)

These essays are worth reading not only for exegetes and patristics scholars interested in digital editions of Greek or Syriac, but also for theologians who want to become acquainted with the creation of digital editions and get a glimpse into the specifics of theological corpora. With the help of von Stockhausen's guidelines, theologically relevant works can be transcribed so they can be reused and connected, thus expanding the field of Computational Theology. According to Zahnd's observations, an expansion is urgently needed. The Geneva Reformation historian points out that conventional ways are rarely abandoned in church history research "because the focus is too heavily on traditional methods of reading" (2020, 117). He continues:

> This is regrettable because the digital medium would offer a wealth of possibilities for further forms of consultation – from digital style analyses and topic modeling to the automated evaluation of intertextual references – which would, however, require other, mechanically evaluable approaches to the editorial data.

The 2022 issue of the *Journal of Ethics in Antiquity and Christianity* on the topic of *Distant Reading – Perspectives of a Digital Age* provides several analyses in the patristic (and New Testament) field, which can be used to demonstrate the "rich possibilities of further forms of consultation:"

- J. Heilmann: Antike Ethik aus der Distanz. Computationelle Methoden zur Erforschung der Ethik im Neuen Testament und im antiken Christentum (12–30)
- Ch. A. Nunn: Das Thema patristischer Ethik – Versuch einer Annäherung durch Distanz am Beispiel der Briefe des Augustinus von Hippo (31–51)
- N. Nikki, V. Kaše & Z. Špiclová: The Cultural Evolution of Prototypical Paul in the First Five Centuries: A Distributional Semantic Analysis of Greek Christian Texts (52–76)
- B. Brunner: "wie Chrysostomus schreibet" – Kirchenväterzitate als normative Referenzen für den Umgang mit Trauer in frühneuzeitlichen Funeralschriften (77–99)
- B. Totsche: Chancen und Grenzen der distant reading-Analyse antiker Texte mit Hilfe von MAXQDA (106–115)

The journal issue not only addresses the possibilities and examples of computer applications; it also deals with infrastructural problems and ideological reservations in this regard (see esp. *Digital Humanities – zwischen Rückschritt und Fortschritt. Ein Standpunkt,* Heil [2022, 103 f.]; cf. also Heilmann [2022, 12–14]).[10] Furthermore, anthologies dedicated to a specific topic and expound upon on practices that could be attributed to Computational Theology should also be mentioned, e. g., the use of network analyses for mapping late antique clerics in exile (Hillner 2016, 11–47).

The emphasis on patristic scholarship here might be partly due to my own affiliation with this field of study.[11] However, the focus on ancient history is well-founded, since in the beginnings of computer use in the humanities, from around the 1960s onwards, "text corpora […] from (English) literature and antiquity […] were digitized and prepared for machine processing" (Haber 2011, 12) and, as a result, "ancient history […] has had access to an almost completely digitized corpus of sources for several years now, which in turn has changed the planning and implementation of research projects" (ibid., 152).[12] However, as the Latinist Revellio (2022, 77–79) claims, there are also problems with the reliability and accessibility of digital text corpora (cf. the chapter from C. T. Schroeder in this volume, esp. pp. 384–386).

2.1.3 Computational Theology in Systematic Theology

One of the contributions to the special volume in *Open Theology* 5(1) is dedicated to systematic theology. Robinson (2019, 67) gives a hesitant credence to the relationship between DH and ST:

> It remains the case that systematic theology has found little use for digital humanities in its teaching and research beyond the use of e-learning platforms and electronic library resources common to many humanities disciplines.

Robinson sees relatively few possibilities for the productive use of DH in ST. In the context of the 2021 annual conference of the *Global Network for Digital Theology,* whose contributions were recorded in the journal *Cursor_* (and could almost all be assigned to type 2 of the DH), Wormstädt offers a critique of Robinson's contribution, emphasizing the potential for integrating DH into systematic theology. He thus identifies five "basis operations" of systematic theology (8):

10 Volp (2020, 439–460) provides a comprehensive overview of the relationship between DH and patristics.

11 Pietism research should also be mentioned. For an overview of the DH in this field, see Faull (2021, 14–18).

12 Revellio (2022, 69–74) treats the question of why ancient studies have such an affinity for digital methods.

1. the analysis of faith statements
2. the analysis of theologically relevant texts
3. the applications of theological insights to non-theological problems
4. the writing of texts
5. the construction of a consistent theological system

Wormstädt thinks there is potential in approach 2 and provides examples in support for this approach (9):

> For example, it might be quite useful to know, whether a systematic theologian is inspired by prior texts which they do not mention [cf. the chapter from J. Nantke on intertextuality research in this volume]. In keeping of distant reading approaches, analyses of an author's oeuvre with regard to typical theological phrases or already identified ones, which proved to be typical for them, could reveal patterns, theoretical/dogmatical emphases, or changes over time within these [cf. the chapter by E. Gius on NER in this volume]. Both might further strengthen hermeneutical findings stemming from close reading research or revealing new leads. Given a well performed hermeneutical research basis, and ever more digitally accessible bodies of text, one might even want to compare bigger corpora, looking for example for trends in German systematic theology compared to those in Scandinavia, US-America, or Nigeria [cf. the chapter by A. Lasch on discourse analysis in this volume].

Wormstädt (2021, 10) also identifies potential for approach 4 and refers to the use of AI-generated texts as an interpretive aid and a challenge to systematic theology ("DeepBarth"; cf. the chapter by J. Gröpler, M. Mundorf, and N. Wilder on AI supported text production in higher education in this volume). However, these approaches are just suggestions without a concrete basis. Wormstädt does not contradict Robinson's observation that ST has very few current points of contact with DH. One can see this by the title of his article: *Relationship status: 'It's complicated.' Digital Humanities Methods and Systematic Theology.*

In another sense, however, points of contact can already be seen. In the DH, efforts to provide a theoretical foundation for research practices have been growing in recent years (cf. Kleymann 2023, 8 f.). Systematic theologians can make important contributions to this discourse. Van Oorschot (2021, 143–164) should be mentioned with her contribution on the topic of *Neue Technik – neue Wissenschaft? Wissenschaftstheoretische und -ethische Herausforderungen der Digital Humanities.* Moreover, it is crucial to conduct monitoring of DH practices, particularly with respect to AI research (cf. Dobson (2019) 29: "The digital humanities need the hermeneutics of suspicion, especially as it applies to methodological choices and interpretations."). Theological ethicists can play an important role here, especially in collaboration with computer scientists. Geldhauser & Diebel-Fischer (2024, 1–6), for example, investigate whether the concepts of fairness and diversity can be implemented mathematically and taught to AI, which proves to be highly problematic.[13]

2.1.4 Computational Theology in Religious Studies

Anyone wanting to learn more about the practices of Computational Theology in Religious Studies is encouraged to consult the six-volume comprehensive treatment found in *Introductions to Digital Humanities – Religion:*

- Vol. 1: D. Veidlinger (Eds.), Digital Humanities and Buddhism (2019)
- Vol. 2: Ch. D. Cantwell & K. Petersen (Eds.), Digital Humanities and Research Methods in Religious Studies (2021)
- Vol. 3: C. van Lit & J. J. Morris (Eds.), Digital Humanities and Religions in Asia (2024)
- Vol. 4: T. Hutchings & C. Clivaz (Eds.), Digital Humanities and Christianity (2021)
- Vol. 5: C. B. Anderson (Eds.), Digital Humanities and Libraries and Archives in Religious Studies (2022)
- Vol. 6: E. S. Clark & R. McBride Lindsey (Eds.), Digital Humanities and Material Religion (2022)

Volume two contains several projects and studies that focus on DH and religiously related subjects. Koch (2020, 152–158) gives a concise overview of the field.

2.1.5 Computational Theology in Practical Theology

Karcher (2020, 132–142) claims there is a tendency in practical theology to do a "theology of the digital" instead of a "theology with the digital." The compendium *Theologie und Digitalität* (Beck et al. 2021), where they map out the entire field of practical theological reflection on the digital is a paradigmatic example of this interpretation.

Nevertheless, a "theology with the digital" can also be found in practical theology, as Karcher (2020, 138–141) provides evidence of this phenomenon based on only three studies. Consequently, his conclusion remains cautious:

13 Cf. also Görder & Zeyher-Quattlender (2019) on the use of data in business, social welfare, and the church from an ethical perspective. See also Puzio, Kunkel & Klinge (2023) on theological approaches to technology and AI.

> Whether practical theology will prove to be a part of DH or a theology of the digital will be decided in the future. Whether the two will continue to develop alongside each other or even overlap also remains to be seen because, thus far, theological projects have hardly been represented in the German-speaking DH community, continue to concentrate on databases and digital editions, or simply exist in a niche with linguistics, communication, history, or religious studies.

Even still, little has changed in Karcher's predictions about the fields. The CONTOC study, for example, which examined the online activities of churches in times of Covid, does not contain any computer text analyses, despite the appropriateness of methods like topic modeling for digital discourse analysis (on the design of the study, see Beck et al. [2023, 25–31]). However, other recent volumes contain at least traces of Computational Theology, which could be further developed. For example, Wünsche et al. (2023, 244–246) use the open access network analysis tool *Gephi* in the volume *Religion auf Instagram* to visualize the linking of around forty central topics with which Muslim Instagram influencers are concerned.[14] In the same volume, Novak et al. (2023, 270–274) present the *YouBeOn Map,*[15] which was created in collaboration with the *Austrian Centre for Digital Humanities* and maps the online and offline environments of religious young people from the greater Vienna area. To broaden the scope of practical theology, digital pastoral care should also be addressed, even if an anthology does not yet exist. Blackstein (2023, 172–183) discusses the use of AI in pastoral care, names specific tools and encourages religious scholars not to reject them outright as dangerous, but to use them creatively in the future as an accompanying (!) instrument "to reach people who may have difficulties seeking out traditional pastoral care services" (181).

2.1.6 Computational Theology and Theological Education

Garcés & Heilmann (2017, 47–49) recommend teaching minimum IT standards in the curriculum to be able to engage productively with digital methods and use them for exegetical research. Three further contributions in the same volume provide examples of how digital teaching can be successfully integrated into (here, biblical exegesis) theological studies and the added value that this approach brings compared to purely analog practices:

14 On the use of social network analysis to analyze religious groups, see Campbell & Sheldon (^{2}2022) 75f.; cf. also Roleder (2020).

15 Cf. https://app.youbeon.eu (Accessed: 13 June 2024).

- K. Künzl & F. Wegschneider: Faszination Digital Humanities. Was benötigen Studierende in ihrer bibelwissenschaftlichen Ausbildung? (53–67)
- T. Flemming: Lernen an Handschriften. Studierende als Experten gewinnen (69–79)
- M. Munson: Natural Language Processing (NLP) unterrichten. Ein Bericht aus der akademischen Praxis (87–92)

In 2020, Giercke-Ungermann and Handschuh published the collected volume *Digitale Lehre in der Theologie. Chancen, Risiken und Nebenwirkungen.* However, concrete DH practices are not mentioned. The focus of the volume is on media didactics (e.g., the role of the teacher in digital settings) and ethical considerations relating to virtual teaching. Schöning (2020, 123) explicitly devotes a paragraph to the use of DH, where he warns against using certain methods as an end in themselves, which may restrict students in their working methods, proving rather to be an obstacle to learning:[16]

> This is particularly true in the field of Digital Humanities. If this term is used to describe the processing of questions in the humanities using digital methods, methodological skills can be developed that follow the rules of the tools. However, only when the techniques are applied in a poetic and playful way that extends the rules can they also be approached in a reflexive and analytical-critical manner always "considering their epistemological imbalances and inherent dynamics" and thus achieve higher taxonomy levels of learning. Such a creative expansion of existing actions takes learners seriously as subjects by challenging them at a high level to do something that cannot be achieved through digitalization (cf. also Schöning 2021, 59).

Programming skills would thus have to be taught in educational fields, which does not seem realistic (cf. Garcés & Heilmann 2017, 46: "Basic computer science training cannot be provided as part of the curriculum in biblical studies [and all other theological subjects]"). The functional scope of program curricula is limited; however, learning digital skills and tools offers simple solutions to introduce students to working with digital methods so that they can get an idea of what kind of research could

16 The danger that computer environments restrict research questions is referred to within DH as "surface theory" or "interface theory" and goes back to Johanna Drucker. See Drucker (2011, 9). Cf. also Berry & Fagerjord (2017, 127): "The surface, or interactional layer, of the digital is hugely important for providing the foundations through which we interact with these technologies. Not only are the interfaces responsive to our questioning via queries, searches, navigation and so forth, they are also designed, increasingly, to be both intuitive, intelligent and contextual, and aesthetic, stylish and pleasant. Modern interfaces often attempt not only to guess our intentions but also to invite extended use and shape the direction of our minds' travel."

be performed with them.[17] Therefore, in my opinion Risam & Gil's (2022, 9) *Minimal Computing* approach is most apt for teaching and education:

> By giving up what might ultimately be prettier or more elegant, in the context of teaching, she instead focused on technologies that help students gain confidence in their digital literacy skills and have small wins that might later encourage them to develop a stronger technical skill set.

Problem solving strategies are practiced in this sense. These strategies are in accordance with what Harrich & Hiepel (2021, 91) call "Computational Thinking" referring to a concept of Jeanette M. Wing:

> The concept of *Computational Thinking* encourages theology to work in this direction. Digital tools in all fields that can be easily used, e.g., when it comes to the processing of raw data, allow research problems to be solved more efficiently that would have been almost impossible to solve without such tools, or would only have been possible with a large consumption of resources.

The essay is in the edited volume by Burke, Hiepel, Niggemeier, and Zimmermann on *Theologiestudium im digitalen Zeitalter,* which offers more possibilities for Computational Theology. For example, Hiepel & Niggemeier (2021, 201–214) discuss how the potential of digital tools could be used in all biblical disciplines. Brockmann et al. (2021, 215–231) show how such tools could contribute analogically to the field of church history. Lüstraeten (2021, 303–317) offers an overview of digitization in liturgical studies.

The 2023 handout *Digitalisierung der Fachbereiche. Theologie und verwandte Disziplinen* describes the current *state of the art.* The authors rightly claim (22):

17 This should not be misunderstood as a plea to use a *black box.* The programs – whether in research or teaching – should contain technical documentation and be openly accessible. Cf. Pirker (2021, 194): "The software and hardware platforms with their specific economic interests in a competitive education market are independent actors with divergent, often not publicly visible interests that are not limited to profit maximization. The networks that enter a relationship and have diverging interests are generally not visible to users as actors. Data-based applications generate both observation possibilities and data links to an extent that has not yet been reflected upon in the world of education. If one reflects on this responsibly in terms of religious education, this means that the selection of tools and didactic paths must be accompanied by a fundamental orientation towards open educational resources – open access and open source, platform-independent offers that are created in open communicating networks, strict conformity with EU data protection directives, consideration of inclusion and diversity – to name just a few criteria." See also the article on research data management by J. Apel in this volume (p. 396).

> The aim of digitalization should be a new culture of teaching and learning, research, and work for all university members, rather than an improvised electrification of the old or mere technical equipment for its own sake. In every respect, this is also a question of openness to the issue of successful teaching in the present and future.

However, the focus is not on specific methods or practices, but on progressive didactic scenarios such as the *flipped classroom* format. The field of Computational Theology is certainly part of this progress, yet also presenting a challenge given the technical skills required (21):

> The general trend towards dependency on third-party funding can also be seen in theology. The funding includes the relation of research fields with a specific application to the present being more likely to receive funding than historically focused fields. Although source research, text research, archaeology, editions, or historical research within theology can also have low potential for third-party funding in digital-related research practices. One problem obviously lies in the fact that the historical theological disciplines would have to develop their own skills to contribute to this field.

The *Compendium* contributes to this set of problems in several ways. On the one hand, the volume can help build appropriate skills, and on the other hand, the contributions make it clear that subjects with contemporary relevance sometimes face even greater problems (e.g., regarding licenses) and that theological researchers beyond the historical subjects could also make such skills fruitful in their field.

2.2 Current German-speaking Locations for Computational Theology

The following (not exhaustive, for more information see the future *living handbook*) presents current projects located at various places in the German-speaking field of Computational Theology. The creation of digital editions (or databases), which exist almost everywhere, is not considered here;[18] instead, cf. the Appendix.

18 For example, Mainz is not specifically mentioned, although the *Institute of European History* (IEG) is home to an entire center for the creation and processing of theologically relevant digital editions. The in-house DH Lab, which provides a digital research infrastructure, is of great advantage here. See https://www.ieg-mainz.de/forschung/dh-lab (Accessed: 15 June 2024). The IEG is the home of some church historians of Reformation history and the early modern period who belong to the DH in some way and can thus also be assigned to Computational Theology. Some include Irene Dingel, Christopher Voigt-Goy, Benedikt Brunner, and Markus Müller (now working on a digital intertextuality project on the Mainz preacher Johann Wild at the University of Wuppertal).

2.2.1 WGTh [Bonn and Göttingen]

The *Wissenschaftliche Gesellschaft für Theologie e. V.* (WGTh) has recently developed forms of Computational Theology, as can be seen from the three-year funding of a new project group, which was applied for by the church historians Aneke Dornbusch (Bonn), Claudia Kampmann (Bonn), and Dorothee Schenk (Göttingen). This project group will focus on network research in church history and theology by exploring the diverse applications of social network analysis through two documented conferences and at least two workshops.[19]

2.2.2 Bochum

Religious studies scholars interested in Computational Theology will find a lot of opportunity for digital research in Bochum. The interdisciplinary Center for Religious Studies (CERES) has its own DH department (DH@CERES) with numerous projects and events that combine DH and religious studies research.[20] The digital research infrastructure of the CRC *Metaphors of Religion* deserves special mention. With the *Akita* annotation tool, for example, a common methodological basis is being developed for all sub-projects to enable "comparative research across languages and religious traditions." The DH@CERES is coordinated by Frederik Elwert, thanks to whom religious research can already be found in publications devoted to the DH (cf. Elwert 2021, 172–186). CERES will soon be supported by a junior professorship for Digital Humanities in Religious Studies.

Not only religious, but also homiletics scholars can find interesting crossover with Computational Theology in Bochum. In the BMBF project RUNIP *(Recht und Normen in Predigten. Maschinell unterstützte Analyse von Predigtkorpora im Zeitvergleich)* under the direction of Markus Totzeck, the historical sermon corpus of Friedrich Schleiermacher and contemporary Protestant sermons are being examined by machine to determine how and to what extent normative arguments are made.[21]

19 For a description of the project, see https://www.wgth.de/images/2024/Projektgruppe%20Netzwerke.pdf (Accessed: 15 June 2024).

20 Cf. https://dh.ceres.rub.de (Accessed: 15 June 2024).

21 For a description of the project, see https://runip-projekt.ruhr-uni-bochum.de/index.html (Accessed: 15 June 2024).

2.2.3 Darmstadt, Rostock, and Wismar

The Darmstadt systematic theologian Gotlind Ulshöfer, the Rostock New Testament scholar Soham Al-Suadi, and the Wismar computer scientist Frank Krüger are leading a joint DFG project entitled *GenderVarianten_Revisionen von Genderkonstruktionen in Textüberlieferungen.* In this project, the manuscript tradition of the New Testament is being examined using machine processing for different understandings of gender. The project thus productively combines textual criticism and gender studies while using DH.[22]

Rostock could also make an important contribution to the integration of Computational Theology into theological education. As part of the *DiCaRo* project, Soham Al-Suadi and communications engineer Tobias Weber are leading a sub-project to develop an interdisciplinary and inter-faculty module aimed at promoting data literacy. The theology department is the pilot.[23]

2.2.4 Munich

Since the company *OpenAI* presented a publicly accessible and freely usable Large Language Model (LLM) in November 2022, AI technologies have increasingly permeated everyday life (see the chapter by Gröpler, Mundorf, and Wilder on AI-supported text production in higher education in this volume). New Testament scholar Christoph Heilig is researching how this movement affects theology.[24] Heilig is the lead investigator in a research group at LMU Munich that is looking at the potential of LLMs in narratological research perspectives on biblical texts.[25]

As part of the DFG project *Zeitgeist und Christentum. Die Zeitschrift Christliche Welt als Medium des Kulturprotestantismus,* systematic theologian Marieluise Sonnemeyer is analyzing the concept of crisis in her doctoral project. She uses a wide range of methods of digital discourse analysis for this purpose.

22 See the project page, https://gepris.dfg.de/gepris/projekt/513300936 (Accessed: 15 June 2024).

23 See the project page, https://www.dicaro.uni-rostock.de/teilprojekte/diss-data-literacy (Accessed: 15 June 2024).

24 New Testament scholar Nicole Oesterreich from the University of Leipzig also names the influence of AI on the development of biblical studies as a research focus. Oesterreich is head of the *Corpus Judaeo-Hellenisticum Novi Testamenti* digital project at the Saxon Academy of Sciences, which was launched in January 2024. Cf. https://www.theol.uni-leipzig.de/institut-fuer-neutestamentliche-wissenschaft/forschung/corpus-judaeo-hellenisticum-novi-testamenti-digital (Accessed: 15 June 2024).

25 Cf. https://www.early-christian-narratives.com/post/ai-diversity-and-marginalized-perspective (Accessed: 15 June 2024).

2.2.5 Passau

Under the title *Digital Methods in Theology,* Christian Handschuh (Church History), Bernhard Bleyer, and Stefanie Müller (Theological Ethics) are cooperating with Annette Hautli-Janisz (Computational Rhetoric and Natural Language Processing). The focus of this cooperation is diachronic argument mining, i.e., the analysis of the temporal change of argumentation patterns in certain (theological) discourses. The initial project focuses on *Die katholische Diskussion um den Suizid zwischen 1800 und der Gegenwart.* Through this project, the potential of digital discourse analysis for theological research in the field of historical and systematic theology is evident.[26]

2.2.6 Zurich

The university research focus “Digital Religion(s)” under the direction of the practical theologian Thomas Schlag is particularly relevant here. The aim of this UFSP is to analyze religious players in the digital space. All areas of Digital Theology are considered in the interdisciplinary projects. Computational Theology practices also have an influence, especially in projects involving cooperation with computational linguistics. For example, computational methods are used to investigate religious mourning communication on Twitter (see N. Bodenmann 2023), or to examine argument structures on religious issues in the digital space between respect and intolerance.[27]

2.2.7 Heidelberg

The TheoLab, based in Heidelberg, is a research network at the interface of theology and DH, which forms the infrastructural framework for the creation of this compendium. In addition, various events have been offered since 2019 to promote research in the field of Computational Theology and networking between members of theology and DH. So far, these efforts have included early career research colloquia, workshop reports, and conferences. Further formats are in the planning phase.[28]

26 See the project page, https://www.ktf.uni-passau.de/digital-methods (Accessed: 15 June 2024).
27 See the project page, https://www.digitalreligions.uzh.ch/de/research/externaldynamics/p8_argument_mining_detection_of_extremism_intolerance.html (Accessed: 15 June 2024).
28 For the TheoLab blog, see https://theolab.hypotheses.org (Accessed: 15 June 2024).

2.3 Summary – Theology in the Canon of the Digital Humanities

The above analysis of the field shows that although there are disciplinary and local differences in the extent to which Computational Theology is practiced, activities in this area are continuing to increase. With the *Compendium Computational Theology,* the editors hope that these activities will be expanded and that the place of theology in Sahle's diagram can be self-evident.

References

Al-Suadi, S. (2021). Historical critical exegesis and the reappraisal of the beginnings of Christianity in the digital age. In ead. & G. Ulshöfer (Eds.). *Biblical Texts in Digitalization. Towards Critical Digital Biblical Studies* (pp. 51–78). Tübingen: Tübingen University Press. DOI: https://doi.org/10.15496/publikation-48536 (Accessed: 15 June 2024).

Ead., & Ulshöfer, G. (Eds.). (2021). *Biblical Texts in Digitalization. Towards Critical Digital Biblical Studies.* Tübingen: Tübingen University Press. DOI: https://doi.org/10.15496/publikation-48536 (Accessed: 15 June 2024).

Anderson, C. (2019). Digital Humanities and the Future of Theology, *Cursor_ Zeitschrift für explorative Theologie,* 1, 75–103. DOI: https://doi.org/0.17885/heiup.czeth.2019.1.24000 (Accessed: 15 June 2024).

Baillot, A., Tasovac, T., Scholger, W., & Vogeler, G. (2023). *Digital Humanities 2023. Book of Abstracts. Collaboration as Opportunity (DH2023).* Graz: Zenodo. DOI: https://doi.org/10.5281/zenodo.7961822 (Accessed: 15 June 2024).

Beck, W., Nord, I., & Valentin, J. (Eds.). (2021). *Theologie und Digitalität. Ein Kompendium.* Freiburg i. Br.: Herder GmbH.

Berry, D., & Fagerjord, A. (2017). *Digital Humanities. Knowledge and Critique in a Digital Age.* Cambridge/Malden: Polity Press.

Blackstein, A. (2023): *Digitale Seelsorge. Impulse für die Praxis.* Göttingen: Vandenhoeck & Ruprecht.

Bodenmann, N., Frick, Karina, Görlich, N., Gröbel, L., & Siever, C. M. (2023). Adressierungsstrategien in Trauer-Tweets aus linguistischer und theologischer Perspektive, *Blog medioscop, Centre for Historical Mediology, University of Zurich.* URL: https://dlf.uzh.ch/sites/medioscope/2023/02/22/adressierungsstrategien-in-trauer-tweets-aus-linguistischer-und-theologischer-perspektive (Accessed: 15 June 2024).

Brockmann, Th., Daufratshofer, M., & Pfister, M. (2021). Digitale Lehr- und Lernformate in der Kirchengeschichte. In A. Burke, L. Hiepel, V. Niggemeier & B. Zimmermann (Eds.), *Theologiestudium im digitalen Zeitalter* (pp. 215–231). Stuttgart: W. Kohlhammer.

Campbell, H.A., & Sheldon, Z. ([2]2022). Community. In H.A. Campbell & R. Tsuria (Eds.), *Digital Religion. Understanding Religious Practice in Digital Media* (pp. 71–86). Abingdon/New York: Routledge.

Clark, E.S., & McBride Lindsey, R. (2022). Digital Materiality. Possibilities of Interpretive Frameworks in the Academic Study of Religion. In eaed. (Eds.), *Digital Humanities and Material Religion. An Introduction* (pp. 1–21). Berlin/Boston: De Gruyter [= *Introductions to Digital Humanities. Religion*, 6].

Clivaz, C. (2020). New Testament and Digital Humanities. In K. Heyden & B. Schröder (Eds.), *Theologie im digitalen Raum* (pp. 98–104). Gütersloh: Gütersloher Verlagshaus [= *Verkündigung und Forschung*, 65(2)].

Collinet, B.J. (2021). Alles auf einen Klick. Digitale Bibelforschung in ihren Chancen und Grenzen erschließen. In M. Datterl, W. Guggenberg & C. Paganini (Eds.), *Digitalisierung – Religion – Gesellschaft* (pp. 147–164). Innsbruck: Innsbruck University Press [= *Theologische Trends*, 31]. DOI: https://doi.org/10.15203/99106-048-2-11 (Accessed: 15 June 2024).

Czachesz, I. (2022). Network Science in Biblical Studies. Introduction, *Annali Di Storia Dell'Esegesi*, 39(1), 9–26. URL: https://asejournal.net/wp-content/uploads/2022/08/czachesz-network-science-in-biblical-studies-2022.pdf (Accessed: 15 June 2024).

DiF-AG Theologie (2023). Handreichung Digitalisierung der Fachbereiche. Theologie und verwandte Disziplinen, *Hochschulforum Digitalisierung. Arbeitspapier*, 71, 1–38. URL: https://hochschulforumdigitalisierung.de/wp-content/uploads/2023/09/Handreichung-dif-Theologie.pdf (Accessed: 15 June 2024).

Dobson, J.E. (2019). *Critical Digital Humanities. The Search for a Methodology.* Urbana/Chicago/Springfield: University of Illinois Press [= *Topics in the Digital Humanities*].

Drucker, J. (2011). Humanities Approaches to Interface Theory, *Culture Machine*, 12, 1–20. URL: https://culturemachine.net/wp-content/uploads/2019/01/3-Humanities-434-885-1-PB.pdf (Accessed: 15 June 2024).

Ensslin, A. (2021). Video Games in/as Digital Humanities? Corpora, Code and Critical Co-Design. In *YouTube-Kanal der Universität Leipzig* (Published: 22 December 2021). URL: https://youtu.be/ou4kR7FpmOM?si=rNInRY5vujsDb5-A (Accessed: 15 June 2024).

Faull, K. (2021). Digital Humanities. In W. Breul (Ed.), *Pietismus Handbuch* (pp. 11–18). Tübingen: Mohr Siebeck. DOI: https://doi.org/10.1628/978-3-16-159652-0 (Accessed: 15 June 2024).

Feldtkeller, A. (2006). Religions- und Missionswissenschaft. Was den Unterschied ausmacht für das Gesamtprojekt Theologie. In I. Dalferth (Ed.), *Eine Wissenschaft oder viele? Die Einheit evangelischer Theologie in der Sicht ihrer Disziplinen* (pp. 121–139). Leipzig: Evangelische Verlagsanstalt [= *Forum Theologische Literaturzeitung*, 17].

Fischer, F. (2019). Digital Classical Philology and the Critical Apparatus. In M. Berti (Ed.), *Digital Classical Philology. Ancient Greek and Latin in the*

Digital Revolution (pp. 203–219). Berlin/Boston: De Gruyter Saur [= *Age of Access? Grundfragen der Informationsgesellschaft,* 10]. DOI: https://doi.org/10.1515/9783110599572 (Accessed: 15 June 2024).

Garcés, J, & Heilmann, J. (2017). Digital Humanities und Exegese. Erträge, Potentiale, Grenzen und hochschuldidaktische Perspektiven, *Forum Exegese und Hochschuldidaktik. Verstehen von Anfang an,* 2(2), 29–52.

Geldhauser, C., & Diebel-Fischer, H. (2024). Is diverse and inclusive AI trapped in the gap between reality and algorithmizability? In *Proceedings of Machine Learning Research,* 233 (1–6). Tromsø: Northern Lights Deep Learning Conference. URL: https://proceedings.mlr.press/v233/geldhauser24a/geldhauser24a.pdf (Accessed: 15 June 2024).

Görder, B., & Zeyher-Quattlender, J. (Eds.). (2019). *Daten als Rohstoff. Die Nutzung von Daten in Wirtschaft, Diakonie und Kirche aus ethischer Perspektive.* Berlin: LIT Verlag [= *LLG Leiten. Lenken. Gestalten,* 40].

Haber, P. (2011). *Digital Past. Geschichtswissenschaft im digitalen Zeitalter.* Munich: Oldenbourg Verlag.

Haberer, J. (2015). *Digitale Theologie. Gott und die Medienrevolution der Gegenwart.* Munich: Kösel.

Harrich, S., & Hiepel, L. (2021). Computational Thinking – ein Ansatz für die Theologie? In A. Burke, L. Hiepel, V. Niggemeier & B. Zimmermann (Eds.), *Theologiestudium im digitalen Zeitalter* (pp. 87–96). Stuttgart: W. Kohlhammer.

Heil, U. (2022). Digital Humanities – zwischen Fortschritt und Rückschritt. Ein Standpunkt, *JEAC* 4, 101–104.

Heilmann, J. (2022). Antike Ethik aus der Distanz. Computationelle Methoden zur Erforschung der Ethik im Neuen Testament und im antiken Christentum, *JEAC* 4, 12–30.

Hiepel, L., & Niggemeier, V. (2021). Nicht auf Sand gebaut. Bibelwissenschaftliche Lehre im digitalen Zeitalter. In A. Burke, L. Hiepel, V. Niggemeier & B. Zimmermann (Eds.), *Theologiestudium im digitalen Zeitalter* (pp. 201–214). Stuttgart: W. Kohlhammer.

Hillner, J. (2016). Approaches to Clerical Exile in Late Antiquity. Strategies, Experiences, Memories and Social Networks. In ead., J. Ulrich & J. Engberg (Eds.), *Clerical Exile in Late Antiquity* (pp. 11–43). Frankfurt a.M.: Peter Lang [= *Early Christianity in the Context of Antiquity,* 17].

Hutchings, T. (2015). Digital Humanities and the Study of Religion. In P. Svensson & D. Goldberg (Eds.), *Between Humanities and the Digital* (pp. 283–294). Cambridge, Mass.: The MIT Press.

Id., & Clivaz, C. (2021). Introduction. In eid. (Eds.), *Digital Humanities and Christianity. An Introduction* (pp. 1–17). Berlin/Boston: De Gruyter [= *Introductions to Digital Humanities. Religion,* 4].

Karcher, S. (2020). Praktische Theologie und Digital Humanities. In K. Heyden & B. Schröder (Eds.), *Theologie im digitalen Raum* (pp. 132–142). Gütersloh: Gütersloher Verlagshaus [= *Verkündigung und Forschung,* 65(2)].

Kleymann, R. (2023). Theorie. In AG Digital Humanities Theorie. Association for Digital Humanities in the German Speaking Areas (Ed.), *Begriffe der Digital Humanities. Ein diskursives Glossar.* Wolfenbüttel: Herzog August Bibliothek [= Zeitschrift für digitale Geisteswissenschaften. Working Papers, 2]. DOI: https://doi.org/10.17175/wp_2023_013 (Accessed: 15 June 2024).

Koch, A. (2020). Religionswissenschaft und digital humanities. In K. Heyden & B. Schröder (Eds.), *Theologie im digitalen Raum* (pp. 152–158). Gütersloh: Gütersloher Verlagshaus [= *Verkündigung und Forschung,* 65(2)].

Lüstraten, M. (2021). Sinnenfällig und schwerfällig. Liturgie und Liturgiewissenschaft und die Herausforderungen der Digitalität. In A. Burke, L. Hiepel, V. Niggemeier & B. Zimmermann (Eds.), *Theologiestudium im digitalen Zeitalter* (pp. 303–317). Stuttgart: W. Kohlhammer.

Moenikes, A. (1997). Zum Verhältnis zwischen Religionswissenschaft und Theologie, *Zeitschrift für Religions- und Geistesgeschichte,* 49(3), 193–207. URL: https://www.jstor.org/stable/23899600 (Accessed: 15 June 2024).

Novak, Ch., Mattes, A., Haselbacher, M., & Limacher, K. (2023). Digital, mobil und gläubig. On- und offline Lebenswelten religiöser Jugendlicher. In V. Pirker & P. Paschke (Eds.), *Religion auf Instagram. Analysen und Perspektiven* (pp. 257–275). Freiburg i.Br.: Herder.

Phillips, P., Schiefelbein-Guerrero, K., & Kurlberg, J. (2019). Defining Digital Theology. Digital Humanities, Digital Religion and the Particular Work of the CODEC Research Centre and Network, *Open Theology,* 5(1), 29–43. DOI: https://doi.org/10.1515/opth-2019-0003 (Accessed: 15 June 2024).

Pirker, V. (2021). Religiöse Bildung im Kontext der Digitalität. Ein kritisch-konstruktiver Blick auf die Zukunftsrelevanz. In N. Brieden, H. Mendl, O. Reis & H. Roose (Eds.), *Digitale Praktiken* (pp. 189–199). Babenhausen: Verlag Ludwig Sauter [= *Jahrbuch für konstruktivistische Religionsdidaktik,* 12].

Puzio, A., Kunkel, N., & Klinge, H. (Eds.). (2023). *Alexa, wie hast du's mit der Religion? Theologische Zugänge zu Technik und Künstlicher Intelligenz.* Darmstadt: wbg Academic [= *Theologie und Künstliche Intelligenz,* 1].

Revellio, M. (2022). *Zitate der Aeneis in den Briefen des Hieronymus. Eine digitale Intertextualitätsanalyse zur Untersuchung kultureller Transformationsprozesse.* Berlin/Boston: Walter de Gruyter GmbH [= *Philologus. Suppl.* 17].

Risam, R., & Gil, A. (2022). Introduction. The Questions of Minimal Computing, *digital humanities quarterly,* 16(2), 1–34. URL: http://www.digitalhumanities.org/dhq/vol/16/2/000646/000646.html (Accessed: 15 June 2024).

Robinson, M. (2019). Embedded, not Plugged-In. Digital Humanities and Fair Participation in Systematic Theological Research, *Open Theology,* 5(1), 66–79. DOI: https://doi.org/0.1515/opth-2019-0005 (Accessed: 15 June 2024).

Roleder, F. (2020). *Die relationale Gestalt von Kirche. Der Beitrag der Netzwerkforschung zur Kirchentheorie.* Stuttgart: W. Kohlhammer [= *Praktische Theologie heute,* 169].

Roueché, Ch., Cameron, A., & Nelson, J.L. (2023). Prosopography meets the digital. PBW and PASE. In J. Nyhan, G. Rockwell, S. Sinclair & A. Ortolja-Baird (Eds.), *On Making in the Digital Humanities. The Scholarship of Digital Humanities Development in honour of John Bradley* (pp. 51–65). London: UCL Press. DOI: https://doi.org/10.2307/j.ctv2wk727j.7 (Accessed: 15 June 2024).

Sahle, P. (2017). Digital Humanities und die Fächer. Eine schwierige Beziehung?, *Forum Exegese und Hochschuldidaktik. Verstehen von Anfang an,* 2(2), 7–28.

Schlag, Th., Nord, I., Beck, W., Bünker, A., Lämmlin, G., Müller, S., Pock, J., & Rothgangel, M. (Eds.). (2023). *Churches Online in Times of Corona. Die CONTOC-Studie. Empirische Einsichten, Interpretationen und Perspektiven.* Wiesbaden: Springer.

Schnelle, U. (2021). *Einführung in die Evangelische Theologie.* Leipzig: Evangelische Verlagsanstalt.

Schöning, B. (2020). Hochschuldidaktik unter der Bedingung der Kultur der Digitalität. In A. Giercke-Ungermann, & Ch. Handschuh (Eds.), *Digitale Lehre in der Theologie. Chancen, Risiken und Nebenwirkungen* (pp. 113–126). Berlin: LIT Verlag [= *Theologie und Hochschuldidaktik,* 11].

Id. (2021). Funktionieren in der Krise. Digitale Technologien und digitalitierte [sic!] Techniken in der Hochschullehre. In N. Brieden, H. Mendl, O. Reis & H. Roose (Eds.), *Digitale Praktiken* (pp. 48–59). Babenhausen: Verlag Ludwig Sauter [= *Jahrbuch für konstruktivistische Religionsdidaktik,* 12].

von Stockhausen, A. (Ed.). (2020) Themenheft. Patristische Editionen im digitalen Zeitalter. Theorie und Praxis, *ZAC* 24(1), 1–160.

Sutinen, E. & Cooper, A.-P. (2021). *Digital Theology. A Computer Science Perspective.* Bingley: Emerald Publishing Limited.

Turner, D. (2005). Doing Theology in the university. In D.F. Ford, B. Quash & J.M. Soskice (Eds.), *Fields of Faith. Theology and Religious Studies for the Twenty-First Century* (pp. 25–38). Cambridge/New York: Cambridge University Press.

van Oorschot, F. (2021). Neue Technik – neue Wissenschaft? Wissenschaftstheoretische und -ethische Herausforderungen der Digital Humanities. In ead. & B. Held (Eds.), *Digitalisierung: Neue Technik, neue Ethik? Interdisziplinäre Auseinandersetzung mit den Folgen der digitalen Transformation* (pp. 143–164). Heidelberg: heiUP [= *FESTForschung,* 1]. DOI: https://doi.org/10.11588/heibooks.945.c12680 (Accessed: 15 June 2024).

Ead. (2023). *Digitale Theologie und digitale Kirche. Eine Orientierung.* Heidelberg: heiBOOKS [= *FEST kompakt,* 7]. DOI: https://doi.org/10.11588/heibooks.1271 (Accessed: 15 June 2024).

Veidlinger, D. (^{2}2022). Religion. A little birdie told me something about religion. Religion on Twitter. In H.A. Campbell & R. Tsuria (Eds.), *Digital Religion.*

Understanding Religious Practice in Digital Media (pp. 132–140). Abingdon/New York: Routledge.

Volp, U. (2020). *computatoribus utamur!* Herausforderungen der Digital Humanities für die Alte Kirchengeschichte. In C. Kampmann, U. Volp, M. Wallraff & J. Winnebeck (Eds.), *Kirchengeschichte. Historisches Spezialgebiet und/oder theologische Disziplin.* FS Wolfram Kinzig (pp. 439–462). Leipzig: Evangelische Verlagsanstalt.

Wormstädt, K. (2021). Relationship status: "It's complicated". Digital Humanities-Methods and Systematic Theology, *Cursor_ Zeitschrift für explorative Theologie,* 1–14. URL: https://cursor.pubpub.org/pub/tjp5i3g0 (Accessed: 15 June 2024).

Wünsche, K., Hafner, A., & von Richthofen, T. (2023). Social Media-Selbst(re)präsentation von Muslim:innen in Deutschland. Akteur:innen, Themen und Positionierungen zu antimuslimisch-rassistischen Diskursen. In V. Pirker & P. Paschke (Eds.), *Religion auf Instagram. Analysen und Perspektiven* (pp. 231–253). Freiburg i. Br.: Herder.

Zahnd, U. (2020). Netzwerke, historisch und digital. Digital Humanities und die Mittlere und Neue Kirchengeschichte. In K. Heyden & B. Schröder (Eds.), *Theologie im digitalen Raum* (pp. 114–123). Gütersloh: Gütersloher Verlagshaus [= *Verkündigung und Forschung,* 65(2)].

Figure Credits

Fig. 1: Spherical Model of the different dimensions of Digital Theology – created by the author.

Fig. 2: Spheres of the Digital Humanities, Version 4 – created by Patrick Sahle, CC BY 4.0

Appendix. Digital Editions in the Field of Theology

Annette von Stockhausen

Two catalogs of digital editions provide an interdisciplinary overview and also contain digital editions from the field of theology:

A Catalog of Digital Scholarly Editions, curated by P. Sahle. URL: https://www.digitale-edition.de/exist/apps/editions-browser/index.html (Accessed: 15 June 2024).
Catalogue Digital Editions, curated by G. Franzini, technically supported by P. Andorfer & K. Zaytseva. URL: https://dig-ed-cat.acdh.oeaw.ac.at (Accessed: 15 June 2024).

However, since neither catalog is complete, I will list a few examples here that show the very diverse or rather disparate status quo of digital editing in the field of theology:

Codex Sinaiticus. URL: https://www.codexsinaiticus.org (Accessed: 15 June 2024). [Review: Schnöpf, M. (2014). Codex Sinaiticus, *RIDE,* 1, 1–28. DOI: https://doi.org/10.18716/ride.a.1.2 (Accessed: 15 June 2024)].
Scripta Qumranica Electronica. URL: https://sqe.deadseascrolls.org.il (Accessed: 15 June 2024).
Editio Critica Maior des Neuen Testamentes. URL: https://ntg.uni-muenster.de (Accessed: 15 June 2024).
Mark16. URL: https://mark16.sib.swiss (Accessed: 15 June 2024). [Cf. Clivaz, C., Monier, M., & Barda, J. (2021). MARK16 as Virtual Research Environment. Challenges and Oportunities in New Testament Studies. In C. Clivaz & G.V. Allen (Eds.), *Ancient Manuscripts and Virtual Research Environments* (no pag.) [= *Special Issue. Classics@Journal,* 18]. URL: https://classics-at.chs.harvard.edu/classics18-clivaz-monier-barda (Accessed: 15 June 2024)].
Patristisches Textarchiv. URL: https://pta.bbaw.de (Accessed: 15 June 2024).
The Saint Patrick's Confessio Hypertext Stack Project. URL: https://www.confessio.ie (Accessed: 15 June 2024). [Review: Brandenburg, Y. (2020). A Review of Confessio.Ie, or Practical Thoughts on Digital Editing in Classics, *RIDE,* 13, 1–51. DOI: https://doi.org/10.18716/ride.a.13.5 (Accessed: 15 June 2024)].
Scholastic Commentaries and Texts Archive. URL: https://scta.info (Accessed: 15 June 2024).
Corpus Thomisticum. URL: https://www.corpusthomisticum.org (Accessed: 15 June 2024).

Hildegardis Bingensis. Liber epistolarum. URL: https://liberepistolarum.mni.thm.de (Accessed: 15 June 2024).
[Cf. Kuczera, A. (2020). TEI Beyond XML – Digital Scholarly Editions as Provenance Knowledge Graphs. In T. Andrews, F. Diehr, T. Efer, A. Kuczera, & J. van Zundert (Eds.), *Graph Technologies in the Humanities. Proceedings 2020* (pp. 101–123). Wien: GRAPH 2020. URL: https://ceur-ws.org/Vol-3110/paper6.pdf (Accessed: 15 June 2024)].

Guillelmus Autissiodorensis, Summa de officiis ecclesiasticis. URL: https://guillelmus.uni-koeln.de/tcrit/tcrit_prologus (Accessed: 15 June 2024).

Der Österreichische Bibelübersetzer. URL: https://bibeluebersetzer-digital.de (Accessed: 15 June 2024).

Wenzelsbibel Digital. URL: https://edition.onb.ac.at/wenzelsbibel (Accessed: 15 June 2024).

Bullinger Digital. URL: https://www.bullinger-digital.ch (Accessed: 15 June 2024).
[Cf. Ströbel, P.B., Fischer, L., Müller, R., Scheurer, P., Schroffenegger, B., Suter, B., & Volk, M. (2024). Multilingual Workflows in Bullinger Digital. Data Curation for Latin and Early New High German, *Journal of Open Humanities Data*, 10(1), 1–13. DOI: https://doi.org/10.5334/johd.174 (Accessed: 15 June 2024)].

Ein Sermon von Ablass und Gnade. a digital edition. URL: https://editions.mml.ox.ac.uk/editions/ablassgnade6 (Accessed: 15 June 2024).

Kritische Gesamtausgabe der Schriften und Briefe Andreas Bodensteins von Karlstadt. URL: http://dev2.hab.de/apps/edoc/start.html?id=ed000216 (Accessed: 15 June 2024).

Briefe und Akten zur Kirchenpolitik Friedrichs des Weisen und Johanns des Beständigen 1513 bis 1532. URL: https://bakfj.saw-leipzig.de (Accessed: 15 June 2024).

Europäische Religionsfrieden Digital. URL: https://tueditions.ulb.tu-darmstadt.de/v/pa000008-0000 (Accessed: 15 June 2024).

Controversia et confessio. URL: https://www.controversia-et-confessio.de (Accessed: 15 June 2024).

Theologenbriefwechsel im Südwesten des Reichs in der Frühen Neuzeit (1550–1620). URL: https://thbw.hadw-bw.de (Accessed: 15 June 2024).

Bibliothek der Neologie. URL: https://bdn-edition.de/index.html (Accessed: 15 June 2024).

schleiermacher digital. Briefwechsel, Tageskalender, Vorlesungen von Friedrich Schleiermacher. URL: https://schleiermacher-digital.de (Accessed: 15 June 2024).

Die sozinianischen Briefwechsel. URL: https://sozinianer.mni.thm.de (Accessed: 15 June 2024).

Kritische Online-Edition der Nuntiaturberichte Eugenio Pacellis (1917–1929). URL: https://www.pacelli-edition.de (Accessed: 15 June 2024).

III. MULTIMEDIA ACCESS IN THE DIGITAL HUMANITIES

Text Digitization

Daniel Stökl Ben Ezra

https://orcid.org/0000-0001-5668-493X

Abstract Text digitization describes the conversion of digital image data of inscribed objects into machine-readable texts.*

Keywords HTR, OCR, Layout Analysis, Handwriting, Machine Learning, Neuron Networks

1. Introduction

Let us define text digitization as the conversion of digital images of inscribed objects of any kind (manuscripts, inscriptions, cuneiform tablets, prints, etc.) into machine-readable texts. Databases with full-text search options in retroactively digitized prints such as JSTOR (since 1994) or Google Books (since 2004) have fundamentally changed the way research is carried out in all academic disciplines. In the last ten years, progress in automatic document analysis, especially in machine learning, has revolutionized the possibilities for researchers to analyze not only difficult prints of the most important cultural texts, but even historical manuscripts, with computerized means.

Transcription of the letters is not the only level of analysis. Text goes beyond a sequence of letters. Gerard Genette (1982) has highlighted the importance of layout and non-main text sections for the (preliminary) understanding of texts. Layout contains critical information, such as the distinction between title and main text, main text and notes, speakers in dramas, verses in poetry, or text connections in translations or commentaries. The choice of and changes in writing style or typeface, width, register (e.g. normal, italic, slanted), weight, color, and alphabets are also essential information carriers that can significantly deepen the depth of analysis of an analyzed text beyond the simple letter sequence (Beinert 2021). In addition, machine paleography, layout analysis, and codicology are used to evaluate these subtle differences for network analysis, dating, and localization of individual objects. This *big data* provides the traditional auxiliary sciences a completely new meaning.

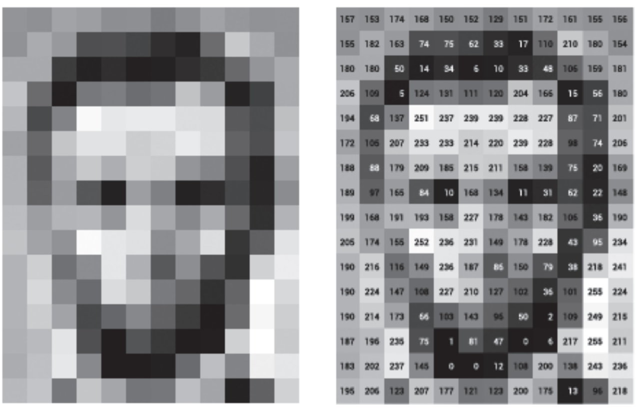
* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

Daniel Stökl Ben Ezra: Text Digitization, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 99–116. DOI: https://doi.org/10.11588/heibooks.1521.c21942

After a brief introduction about the encoding of image and text, neural networks, and existing programs, this chapter treats layout analysis and reading order, computer paleography and text recognition. The analysis is in constant engagement with the discipline known in computer science as "image processing" (Maier et al. 2020).

2. Image Encoding

Computers can only distinguish between the values 0 and 1. Both texts and images must initially be represented as sequences of 0s and 1s. Usually, the image is then transferred with a grid into a table/matrix with rows and columns, in which each cell contains a value for one pixel. High-resolution images have more pixels for the same object surface than low-resolution images. Black and white images are the simplest image format, only knowing 0 or 1 for each pixel, such as foreground (ink) or background (paper). This format leads to the familiar stair steps (Fig. 1), especially at low resolution or high magnification.



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Fig. 1 Abraham Lincoln as a grayscale image with the values on the scale of 0–255

For more nuanced image, one can use grayscale images that allow intermediate levels. 0 still stands for black, but instead of 1, 255 is now used for white, and all numbers in between denote gray values, depending on whether they are closer to white or black. These are so-called 8-bit images, i.e., a combination of eight memory cells (bits) for each pixel ($256 = 2^8$). For finer gradation, one can also use more extensive grayscale and then work with values from 0 to 65,555 (16-bit – 2^{16}) or more. The lowest value is used for black and the highest value for white, and the values in between represent the possible shades of gray.

Color images use a combination of three superimposed grayscale images with an 8-bit scale of 0–255. The color composition (e.g., red-green-blue = RGB) produces the mixed colors for eyes, e.g., violet, brown, orange, or pink. Instead of RGB, other color channels can also be used, e.g., cyan, magenta, and yellow, as in color printers. And, as with grayscale images, a much finer 16-bit scale can be selected instead of an 8-bit scale.

The greater the height and width, resolution or color scale, the more memory an image requires. Image data is often compressed to save space or increase processing speed. Each camera or scanner manufacturer has its own proprietary format *(Raw).* When exported, this format is converted into TIFF, PNG, or JPEG image files. Depending on the compression, Tiff and Png require considerably more storage space than JPEG image files (extension jpg), which are created using a compression algorithm that accepts more loss of information to take up less storage space. This process creates artifacts, which can be easily recognized at high magnification the image looking like bathroom tiles. As long as the image resolution is good (at least 30 pixels for an ‘a’ or ‘א,’ much more for more complex scripts such as Chinese), jpgs are sufficient for layout analysis and text recognition. Automatic paleography or writer identification achieves better results with tiff and png files.

3. Text Coding

In a computer, a text is also stored as sequences of 0s and 1s. In the past, when memory was more expensive, 8 bits were reserved for all variants together for each character, so that computers only knew 256 different values (code points), which could only represent a selection of either Latin, Greek, Cyrillic, or Hebrew, depending on the linguistic or geographical workplace, and was known as the Extended ASCII table. An additional code in the text file indicated which alphabets were meant by the raw numbers in the file. It was thereby possible to work with the local alphabets in Germany, Bulgaria, Israel, or Saudi Arabia, but never with all scripts at the same time. More complex scripts with a total number of characters greater than 256, such as Chinese, were impossible. Furthermore, if the encoding schema was unknown, the only solution was to try out all encodings until the text was displayed legibly. This limitation to 256 different characters was a major challenge for philological work in theology, where Syriac, Arabic, Armenian, Georgian, or Coptic – and sometimes also Akkadian, Egyptian, Ethiopian or Sanskrit – were often used in addition to Greek and Hebrew.

The introduction of the Unicode standard in 1991 has progressed towards solving this problem. Like grayscale images with finer nuances, the memory reserved for each character was initially doubled from 8 bits to 16 bits, allowing 2^{16} = 65,536 different code points in the table. In 2022, a total of 161 fonts could be defined in a single table using this encoding, also known as UTF-8. Even though the introduction of the

Unicode was a major step forward, difficulties for the digital modelling of historical scripts remain, e.g., there is still no Unicode for the Babylonian vocalization of Hebrew texts, and Egyptian or Akkadian have only been partially standardized.

Humanities scholars should be familiarized with the intricacies of Unicode because the basic questions of script encoding that need to be solved are tricky. The present chapter will return to further questions, like reading order in bidirectional texts that mix Hebrew and Latin, or character combination coding, in the text recognition section.

4. Neural Networks

The rapid progress in automatic document analysis in recent years has five main interdependent causes: Hardware (memory, speed), software (neural networks), training data volumes, mass digitization and open-source policies. Processors have become much faster and can process much larger amounts of data simultaneously thanks to increased memory. Hard disk storage and internet data transmission (fiber optics) are also cheaper and of much better quality than 10 years ago. Mass digitization projects of culturally significant manuscript collections, archives, and libraries have led to a flood of image data. The interest of large corporations in processing large amounts of written and oral text (Google Books, YouTube, Netflix, Zoom) has not only improved existing algorithms but also developed new ones. Some of these algorithms are freely available to the public in open-source packages of the most important programming languages (e.g., *pytorch* from Facebook, *TensorFlow* from Google). There are research projects that have published their training data under open licenses, thus enabling others to use them to develop or optimize new algorithms.

Different forms of artificial neural networks are used for almost all stages of automatic document analysis. The basic principle has been known since Rosenblatt in 1958, but it was the above-mentioned constellation of simultaneous progress in hardware, software, data, and open source that led to their success (starting with Jürgen Schmidhuber and Yann LeCun's works in the early 1990s). The common principle is a very complex formula with thousands, millions, or even billions of parameters that are optimized by the computer in a learning process called *training*. The result of a trained network architecture is called a *model* because it mathematically models the problem (Fig. 2).

To a certain extent, artificial neural networks imitate the way brains work. The three most relevant network types at present are *Convolutional Neural Networks* (CNN), *Recurrent Neural Networks* (RNN) and Transformers. The most important common principle is the inclusion of context for each data point. CNNs are particularly interesting for images because a data point (i.e., a pixel) is considered in the context of a rectangle. For example, the computer can learn abstract concepts such as curves of different curvature and lines at different angles and orientations and combinations. RNNs are interesting for sequences such as audio recordings or texts, as they are flexible in learning how much context to include for a certain phenomenon. For longer sequences, *Long-Short Term Memory Neural Networks* (LSTM), are often used. In deep learning, several layers of CNNs and/or RNNs are combined, resulting in complex architectures that require large amounts of data and often a lot of time to train. At the same time, these combinations also deliver excellent results, both for layout analysis and for transcription as well as many other tasks like classification.

Transformers have become commonplace through recent *Large Language Models* (LLMs), such as BERT *(Bidirectional Encoder Representations from Transformers)* and GPT *(Generative Pre-trained Transformer).* With the right configuration and an optimal learning process, they can learn to perform layout analysis simultaneously with transcription, recognize proper names (Named Entity Recognition), translate, or summarize the resulting texts or even answer questions.

When training, a distinction is made between supervised, self-supervised, and unsupervised training processes. In the supervised training process, the computer is shown *questions* with the corresponding *answers* and tries to optimize the parameters of the formula to arrive at the answer from the question step by step. Question/answer pairs depend on the learning objective and can be very different depending on the task: (1) the image of a line of text and the corresponding transcription; (2) the image of a manuscript page and the corresponding polygons of the layout; or (3) the image of a book page and the corresponding print type. Both question and answer are represented as a number (scalar), number sequence (vector), matrix or tensor etc., since computers know nothing else. At the beginning, all parameters are often initialized randomly. After each learning step (question/answer calculation), the distance of the calculated answer to the correct one is measured and the parameters are adjusted so that the next time the computer is confronted with the same or similar question, the calculated answer is closer to the correct one. If the distance gradually diminishes (and it is not always the case), then the model is said to converge. At regular intervals, the computed formula is subjected to a test in which it is presented with question/answer pairs without the computer learning from them, i. e., without changing the formula, but only determining the current precision of the current model. The specialist tries to formulate the network architecture so that that it can achieve the best models with as little training material as possible and in the fastest possible computing time. Training is usually terminated when the user determines from the test results that the computer is not improving any further. Finally, the model that delivered the best results in the comparison is saved.

The training process is not uniform. Consider the following illustration: a spaceship is supposed to fly as far as possible through a complex labyrinth of caves but is only allowed to make turns at a certain angle and must then fly a certain distance in this direction. If the selected distance is too short, the spaceship will crawl through the large opening space without finding one of the entrances to the labyrinth. If the

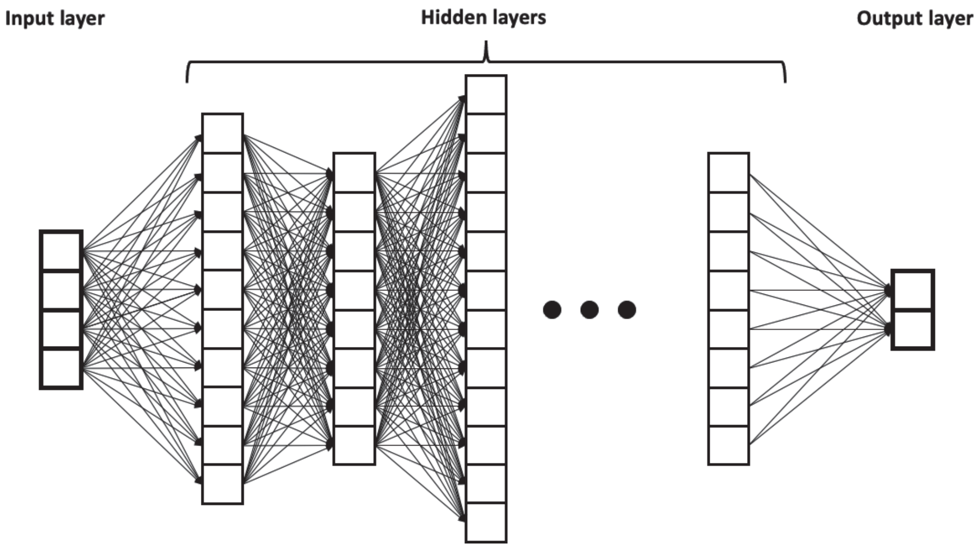


Fig. 2 Example diagram of a neural network with a one-dimensional input layer (with four variables) that leads through a complex network to an output layer with two variables. Each of the lines is a mathematical operation, the parameters of which must be optimized.

distance is too long, the spaceship will not be able to pass through tight turns. Angle and distance are among the *hyperparameters* of the training process, perhaps best compared to the *learning rate.*

In the *self-supervised* training process, the computer automatically calculates the answers from the questions. For example, the computer is shown many images of lines of handwriting, some of which are blacked out, and is asked to suggest a substitute image of what the blacked-out section of the line might have looked like. The model is optimized by analyzing the difference between the computer's suggestion and the original image. In this case, the computer gradually learns the principle of which pixel clusters are usually located between which other pixel clusters. Self-supervised training is sometimes used as pre-training before supervised training because it allows basic principles to be learned beforehand and reduces the amount of training data with manually generated responses.

One potential danger is overfitting. We can compare this to a student who memorizes the answers to the practice questions without understanding the underlying principle. The student can answer the practice questions almost perfectly but is unable to solve unseen questions. The same can happen, for example, if the training corpus is not adapted to the network architecture (too little training material for a question that is too complex) or if the learning rate at which the computer tries to adapt the parameters after each learning process is set too high or too low. Of course, we have only been able to describe a few hyperparameters here.

5. Existing Programs

Until recently, most companies and individual researchers interested in OCR worked with commercial programs like *ABBYY Finereader,* which was very successful with modern English or German texts but could not handle handwritten material or rarer printed texts such as the Syriac alphabets, for which the market seemed too small. Google's widely used open-source program *Tesseract* is only recommended for OCR but not for HTR.[1] Currently, there are several successful programs used for mass digitization of handwritten material.

Since 2016, *Transkribus* has enabled the automatic layout analysis and transcription of written objects, the manual correction of layout analysis and recognition, as well as the training of own transcription models based on the entered data with excellent results via a complex JAVA app or a simplified web app (Kahle et al. 2017, Mühlberger et al. 2019). The program, originally developed in several European research projects, was commercialized in 2019 in the form of a European cooperative.[2] Currently, users pay per page for automatic layout analysis and/or automatic text recognition. The platform and trained models are therefore closed-source. Other commercial programs include *Ocelus* and *Calfa* (Vidal Gorène 2021a).

In open-source, *OCRopus/ocropy,* developed by Thomas Breuel (2008), was a decisive step forward. Although, as the name suggests, it was only developed for OCR, our Paris team, with the help of Marcus Liwicki, has also been using it for handwriting recognition since 2015. However, programming knowledge was a prerequisite for use. There was only a very rudimentary way to enter transcriptions, only for even lines and only very simple segmentation. Since 2018, *eScriptorium* has been developed around Benjamin Kiessling's *Kraken* (Kiessling et al. 2019, Stokes et al. 2021). It is currently the only open-source program for handwriting analysis with an ergonomic user interface for layout and transcription correction as well as text alignment. It can be installed directly on Linux, Mac OS, and Windows computers using WSL *(Windows Subsystem for Linux).* If a team wants to collaborate on the same document(s), a server is required. A GPU with sufficient RAM is needed to train layout or transcription models.

1 FAQ: https://tesseract-ocr.github.io/tessdoc/FAQ.html#can-i-use-tesseract-for-handwriting-recognition (Accessed: 15 June 2024).

2 See https://readcoop.eu/a-short-history-of-transkribus-with-gunter-muhlberger (Accessed: 15 June 2024).

6. Layout Analysis

Computerized layout analysis has two objectives. First, computerized analysis is – thus far – a necessary step before text recognition. Second, the layout contains essential information for the hierarchy, reading order (see below), different text types, differentiation between image and text, etc. Layout analysis is crucial in text comprehension even after text recognition.

Previously, morphological operations were used for layout analysis to recognize different text blocks and lines. Currently, this is accomplished by neural network architectures (Fig. 3) that manage both the segmentation of regions and their division into types (column, header, marginalia, illustration, table, apparatus, etc.) as well as the recognition of lines and their division into types (main text line, interlinear line), and the writing direction (horizontal, vertical, upside down). A segmentation ontology determines which region and line types can be used for which phenomena. Incidentally, zones do not necessarily have to be text regions. Users can also use image segmentation to locate library stamps, coins, illustrations, etc.

There are currently two different approaches to region segmentation. One approach uses principles for object recognition such as traffic lights or signs in self-driving cars (Clérice 2022). This approach works very well for text objects with only rectangular regions that have been digitized in a precise fashion, e.g. without tilting, sheering or rotation. However, problems quickly arise with more complex layouts, e.g., L-shaped regions, or with small rotations.

The other approach is a pixel classifier (Kiessling 2020). All image pixels are assigned to one or more desired types of regions. The pixel cloud is then determined for each region type and one or more polygons are reconstructed. This approach better manages complex layouts or rotated digitized images but has difficulties assigning pixel groups of the same type that are very close to each other to the same polygon. The approach therefore tends to classify two closely spaced main text columns as a single zone.

Line segmentation occurs simultaneously to or after region segmentation (Grüning 2017). There are also two approaches in line segmentation. Either a neural network is trained first to detect the baseline and writing direction of each line and then calculate a polygon that surrounds this baseline so that all ink traces of the characters in this line, including any dots and dashes above or below them, are included. Or the neural network is trained to recognize the line polygon directly and then derive the writing direction.

If the training data is homogeneous and numerous enough, in *kraken/eScriptorium,* very complex segmentation models can be trained with 20 different region and line types (Stökl Ben Ezra 2022b). Training simple specific segmentation models is possible with just a few training pages.

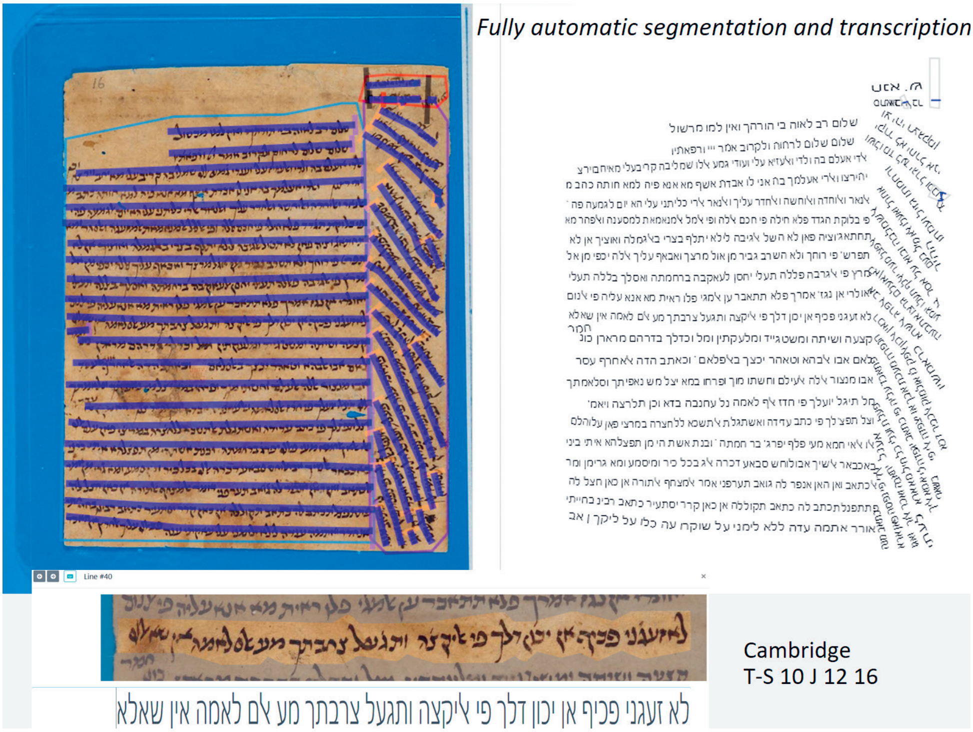


Fig. 3 Fully automated layout segmentation and transcription of the Genizah fragment Cambridge, T.-S. 10 J 12 16 with *eScriptorium*, created in the HTR4PGP project.

7. Reading Order

Different text types diverge in their reading order. There are very different ways to read texts, such as the sections of one or more newspaper pages, a critical edition, a bilingual edition, a table, a manuscript with a basic text or with commentaries, a letter, or postcards. Poetic texts are often written stychographically looking like two columns of a prose text. Reading order analysis is closely related to layout analysis. In *kraken*, layout analysis can currently be trained in a version still under development, the integration of which into *eScriptorium* will take a few more months to design and implement the necessary ergonomic user environment (for the method, see Quirós 2022).

8. Computational Paleography

One area in which neural networks come very close to or even surpass humans is the classification of objects. Computational paleography can be used for a wide variety of purposes. Objects can be classified according to the types of printing or fonts used, schools, or scribes, or even dated or geographically located (Seuret et al. 2021, Droby et al. 2022, Popovic et al. 2021). Banks have been using this technology to verify signatures for a long time. Object classification can also be used to analyze whether a handwriting was written by one or more person. The Friedberg Genizah Project used these techniques early on, where users searching for similar fragments to fragment X, which may originate from the same codex, were suggested other candidates from the approximately 300,000 fragments by the system, which led to a large number of "joins" (Wolf et al. 2010).

9. Text Recognition

Text recognition previously required isolating each character. Modern neural networks, however, are based on entire lines of text and thereby achieve better results, as the context of each data point can be taken into consideration. Furthermore, when recognizing handwriting, isolating letters is difficult as such, and even the correct deciphering of whole words depends on the context of the preceding and following words and characters.

Current state of research does not permit to determine precisely how many lines are needed to train a model, or how often a character must occur in the training data set for the computer to learn it. Learning the lines and characters depends on factors like the complexity of the network, the number of different characters to be learned, legibility and uniformity of the script, contrast, uniformity of orthography and vocabulary, image quality, among others. When creating the training corpus, strict homogeneity of the data must be ensured, especially in the case of collective annotation or the affiliation of different projects. The same phenomenon should always be transcribed in the same way. More training material is required if the scholar wants to teach the computer to resolve abbreviations or to transcribe orthographic variants in a normalized way. Finally, in principle, the computer can also be trained to differentiate between different characters, e.g., between italics, bold and normal, or between different font types.

Usually, a preliminary analysis of the final purpose is worthwile. If visually similar characters are encoded differently, for the computer they are just as apart as visually completely diverse characters are for humans. Do all different types of quotation marks (‘’“”«») really have to be differentiated in exactly the same way, or can some or all of them be standardized? Depending on the project, it may prove advantageous to go into the one or the other direction.

Here it might be helpful to return our attention to Unicode. Some letters appear multiple times in the Unicode table. E.g., there is ‘s’ and ‘ſ’. Using these and other subtle differences, an allographic or hyper-diplomatic transcription can be trained to differentiate between allographs. There are thus fundamental questions that every edition project (but also the Unicode Consortium) must answer: what is the same character, where are differences mandatory, and where should the project give users the freedom to differentiate for themselves?

A circle can be a Latin ‘O’ or ‘o,’ a Greek ‘O’ or ‘o,’ a Hebrew ס (a letter pronounced as ‘s’), or an Arabic ٥ (the sign for ‘5’). What a circle is in a historical document depends on the context. Sometimes the Unicode table distinguishes between different alphabets, but not always. For Latin, English, and German, the same encoding is used for ‘O’. ‘Ö’ is either an ‘O’ combined with a Trema ‘¨’ or a single character ‘Ö’. They cannot be visually distinguished. But if one half of the training data has been created with the combination of ‘O’ + ‘¨’ and the other half with ‘Ö’, then the network becomes confused, since for the computer these are completely different entities. With Greek accents or Hebrew diacritics and vowels, there is also the sequence, because sin, sin-dot, dagesh, and a kamatz can be written in 24 different permutations. Therefore, Unicode includes the possibility of a code point normalization, which (a) either decomposes all characters as much as possible and then puts them in the same order (NFD – decomposed), or (b) combines all characters in the same order (NFC – combined).

Some visually identical characters appear several times in the Unicode table. Latin capital ‘A’ and Greek ‘Α’ (“Alpha”) are visually identical but are represented in the computer by two different encoding points. For the graphically identical letters in Greek and Coptic, there was originally a single common code point for each in the Unicode table. It was not until version 4.1 in 2005 that separate code points were introduced for each.[3] Most of the digits in Persian and Arabic are also identical in appearance, yet they are all encoded twice in the Unicode table, once for Persian, and once for Arabic.

When planning a project, it can pay off to start with simple texts, or even with existing digital texts than with more complex or interesting or non-transcribed texts. Often the best starting point is to train a base model for all documents in a script type capable of transcribing as many scribal hands as well as possible, then apply this base model to a few pages of a new manuscript, correct these pages and then retrain *(fine-tune)* this base model with this new data to optimize it for this particular manuscript. If transcriptions already exist elsewhere, then a large training corpus can be created even faster with text-to-text alignment.

Neural networks trained with purely visual information learn only a very primitive language model linked to the probabilities of a certain character appearing be-

3 See https://en.wikipedia.org/wiki/Greek_and_Coptic (Accessed: 15 June 2024).

tween two or more other characters. However, a more complex language model can also be added during training or as a separate post-OCR step. However, doing so involves the risk of hypercorrection in the case of historically variable orthographies, especially when using language models trained on modern texts.

There are now many transcription projects involving more and more languages and scripts. Perhaps the largest project relevant to theology to date is the ERC Synergy project MiDRASH, which has begun in October 2023 and aims to analyze and transcribe a large part of the approximately 100,000 digital Hebrew manuscripts collected in the KTIV project of the National Library of Israel. Many projects are now uploading their data to the *HTRUnited* catalog.[4] In addition to the projects and publications already mentioned, there are also patristics projects on Greek texts in Berlin (von Stockhausen) and on Coptic texts in Berlin (Lincke 2019, 2021), Oklahoma (Schroeder), and Tokyo (Miyagawa 2018, 2019, 2021). Much work on Armenian and Georgian manuscripts and prints is being done on CALFA (Vidal Gorène 2021). For other data sets, see also Nikolaidou et al. 2022. Bullinger's correspondence is being analyzed in a project led by Tobias Hodel (Scius-Bertrand 2023, Ströbel 2023).

10. Text2Image Alignment

Text2image alignment is indispensable for paleographic studies as well as for digital editions. This type of alignment calculates the approximate regions for each letter and word in a transcription line. With certain neural networks, this approximation is part of the automatic transcription. However, this information is lost when the automatic transcription is corrected manually. With text-image alignment the approximate positions of the letters and words can be recalculated retrospectively. There are current projects underway on the Books of Hours by Dominique Stutzmann (Hazem 2020), on Qumran (Stökl Ben Ezra et al. 2020), on rabbinic texts (Stökl Ben Ezra, in press), and on the Hebrew Bible (Bambaci et al. 2023, Stökl Ben Ezra et al. 2021).

11. Text2Text Alignment

A practical method of using existing high-quality electronic texts for the creation of training data is *text2text* alignment. After correcting an automatic layout analysis, the best existing text recognition model is applied. The computer then calculates how best to align the electronic text with the faulty automatic transcription and swaps the latter with the former. If there are no line or page breaks, then the computer inserts

4 See https://htr-united.github.io (Accessed: 15 June 2024).

them.[5] The *Text2Image* tool in *Transkribus* works in a similar way. The further connection with entire text corpora will allow the automatic creation of training material (Smith 2023).

12. From HTR Platform to Edition

The way from an automatic transcription to a digital edition is not (yet) clear; there are still stumbling blocks in the way. In editions, interlinear and marginal improvements are marked with brackets as additions but integrated into the running text. In transcription platforms, on the other hand, these are in separate lines. A continuous text is essential for the use of collation programs for critical editions of texts with several manuscripts. A pipeline for Hebrew texts promises initial approaches to solutions (Stökl Ben Ezra 2022a).

Literature Cited

Bambaci, L., Stökl Ben Ezra, D. (exp. 2024). Enhancing HTR of Historical Texts through Scholarly Editions. A Case Study from an Ancient Collation of the Hebrew Bible. In *Computational Humanities Research Conference. CEUR Workshop Proceedings* 2023 (pp. 554–576). Paris: Computational Humanities Research. URL: https://ceur-ws.org/Vol-3558/paper6310.pdf (Accessed: 15 June 2024).

Beinert, W. (2021). Schriftstil. In Id. (Ed.), *Das Lexikon der Typografie.* URL: https://www.typolexikon.de/schriftstil (Accessed: 15 June 2024).

Breuel, Th. (2008). The OCRopus open source OCR System. In B. A. Yanikoglu & K. Berkner (Eds.), *Document Recognition and Retrieval XV, part of the IST-SPIE Electronic Imaging Symposium,* San Jose, CA, USA, January 29–31. DOI: https://doi.org/10.1117/12.783598 (Accessed: 15 June 2024).

Droby, A., Irina, R., Vasyutinsky-Shapira, D., Kurar-Barakat, B., & El-Sana, J. (2022). Digital Hebrew Paleography. Script Types and Modes, *Journal of Imaging,* 8(5.143), 1–22. DOI: https://doi.org/10.3390/jimaging8050143 (Accessed: 15 June 2024).

Genette, G. (1982). *Palimpsestes. La littérature au second degré.* Paris: Éditions du Seuil.

Grüning, T., Leifert, G., Strauß, T., Michael, J., & Labahn, R. (2019). A two-stage method for text line detection in historical documents, *International Journal of Document Analysis and Recognition,* 22(3), 285–302.

5 See https://github.com/dasmiq/passim (Accessed: 15 June 2024).

Hazem, A., Daille, B., Kermorvant, Ch., Stutzmann, D., Bonhomme, M.-L., Maarand, M., & Boillet, M. (2020). Books of Hours. The First Liturgical Data Set for Text Segmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 776–784). Marseille: European Language Resources Association. URL: https://aclanthology.org/2020.lrec-1.97 (Accessed: 15 June 2024).

Kahle, Ph., Colutto, S., Hackl, G., & Mühlberger, G. (2017). Transkribus. A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *14th IAPR International Conference on Document Analysis and Recognition* (pp. 19–24). Kyoto: IEEE Xplore. DOI: https://doi.org/10.1109/ICDAR.2017.307 (Accessed: 15 June 2024).

Kiessling, B., Tissot, R., Stokes, P.A., & Stökl Ben Ezra, D. (2019). eScriptorium. An Open Source Platform for Historical Document Analysis. In *International Conference on Document Analysis and Recognition Workshops (ICDARW)* (pp. 19–24). Sydney: IEEE Xplore. DOI: https://doi.org/10.1109/ICDARW.2019.10032 (Accessed: 15 June 2024).

Kiessling, B. (2020). A Modular Region and Text Line Layout Analysis System. In *17th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 313–318). Dortmund: IEEE Xplore. DOI: https://doi.org/10.1109/ICFHR2020.2020.00064 (Accessed: 15 June 2024).

Lincke, E.-S., Bulert, K., & Büchler, M. (2019). Optical Character Recognition for Coptic fonts. A multi-source approach for scholarly editions. In *DATeCH2019 – Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage* (pp. 87–91). New York: Association for Computing Machinery. DOI: https://doi.org/10.1145/3322905.3322931 (Accessed: 15 June 2024).

Lincke, E.-S. (2021). The State of the affairs in optical character recognition (ocr) for Coptic. In C.G. Zamacona & J. Ortiz-García (Eds.), *Handbook of Digital Egyptology. Texts* (pp. 139–164). Alcalá de Henares: Universidad de Alcalá.

Maier, A., Christlein, V., Breininger, K., Würfl, T., Unberath, M., & Riess, Ch. (2021). Bildanalyse. In G. Görz, U. Schmid & T. Braun (Eds.), *Handbuch der Künstlichen Intelligenz* (pp. 673–712). Berlin/Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110659948-015 (Accessed: 15 June 2024).

Miyagawa, S., Bulert, K., Büchler, M., & Behlmer, H. (2019). Optical character recognition of typeset Coptic text with neural networks, *Digital Scholarship in the Humanities,* 34(1), 35–41. DOI: https://doi.org/10.1093/llc/fqz023 (Accessed: 15 June 2024).

Miyagawa, S., Zeldes, A., Büchler, M., Behlmer, H., & Griffitts, T. (2018). Building Linguistically and Intertextually Tagged Coptic Corpora with Open Source Tools. In Suzuki, Ch. (Ed.), *Proceedings of the 8th Conference of Japanese Association for Digital Humanities. Leveraging Open Data.* (pp. 139–141). Tokyo: Center for Open Data in the Humanities.

Miyagawa, S. (2020). Digitization of Coptic Manuscripts and Digital Humanities. Tools and Methods for Coptic Studies, *The International Journal of Levant Studies,* 2, 29–61.

Mühlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, Ph., Kallio, M., Kaplan, F., Kleber, F., Labahn, R., Lang, E.M., Laube, S., Leifert, G., Louloudis, G., McNicholl, R., Jean-Meunier, L., Michael, J., Mühlbauer, E., Philipp, N., Pratikakis, I., Puigcerver Pérez, J., Putz, H., Retsinas, G., Romero, V., Sablatnig, R., Joan-Sánchez, A., Schofield, Ph., Sfikas, G., Sieber, Ch., Stamatopoulos, N., Strauß, T., Terbul, T., Toselli, A.H., Ulreich, B., Villegas, M., Vidal, E., Walcher, J., Weidemann, M., Wurster, H., & Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition, *Journal of Documentation,* 75(5), 954–976.

Nikolaidou, K., Seuret, M., Mokayed, H., Liwicki, M. (2022). A survey of historical document image datasets, *International Journal of Document Analysis and Recognition,* 25(4), 305–338.

Quirós, L., & Vidal, E. (2022). Reading order detection on handwritten documents, *Neural Computer Applications* 34(12), 9593–9611.

Rosenblatt, F. (1958). The perceptron. A probabilistic model for information storage and organization in the brain, *Psychological Review,* 65(6), 386–408. DOI: https://doi.org/10.1037/h0042519 (Accessed: 15 June 2024).

Scius-Bertrand, A., Ströbel, Ph., Volk, M., Hodel, T., & Fischer, A. (2023). The Bullinger Dataset. A Writer Adaptation Challenge. In *ICDAR 2023. Document Analysis and Recognition. Conference Proceedings,* 1 (pp. 397–410). San José: Springer. DOI: https://doi.org/10.1007/978-3-031-41676-7_23 (Accessed: 15 June 2024).

Seuret, M., Nicolau, A., Rodríguez, D.-S., Weichselbaumer, N., Stutzmann, D., Mayr, M., Maier, A., & Christlein, V. (2021). ICDAR 2021. Competition on Historical Document Classification. In Lladós, J., Lopresti, D., Seiichi, U. (Eds.), *Document Analysis and Rcognition. ICDAR 2021* (pp. 618–634). Lausanne: Springer. DOI: https://doi.org/10.1007/978-3-030-86337-1_41 (Accessed: 15 June 2024).

Smith, D., Murel, J., Parkes-Allen, J., & Miller, M.T. (2023). Automatic Collation for Diversifying Corpora. Commonly Copied Texts as Distant Supervision for Handwritten Text Recognition. In *Computational Humanities Research Conference. CEUR Workshop Proceedings* 2023 (pp. 206–221). Paris: Computational Humanities Research. URL: https://ceur-ws.org/Vol-3558/paper1708.pdf (Accessed: 15 June 2024).

Stokes, P., Kiessling, B., Tissot, R., Gargem E., & Stökl Ben Ezra, D. (2021). The eScriptorium VRE for Manuscript Cultures. In C. Clivaz, & G.V. Allen (Eds.), *Ancient Manuscripts and Virtual Research Environments* (no. pag) [= *Classics@Journal* 18(1)]. URL: https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem (Accessed: 15 June 2024).

Stökl Ben Ezra, D., Brown-DeVost, B., Dershowitz, N., Pechorin, A., & Kiessling, B. (2020). Transcription Alignment for Highly Fragmentary Historical Manuscripts. The Dead Sea Scrolls. In *International Conference on Frontiers in Handwriting Recognition* (pp. 361–366). Dortmund: IEEE Xplore. DOI: https://doi.org/10.1109/ICFHR2020.2020.00072 (Accessed: 15 June 2024).

Stökl Ben Ezra, D., & Lapin, H. (in print). From HTR to Digital Critical Scholarly Edition. Reflexions on the Use of Machine Learning. Computational and Digital Humanities in the Sofer Mahir Project. In U. Henny-Krahmer et al. (Eds.), *Machine Learning and Data Mining for Digital Scholarly Editions.* Norderstedt: Books on Demand [= *SIDE*].

Stökl Ben Ezra, D., Lapin, H., Brown DeVost, B., & Jablonski, P. (2022a). HTR2CritEd. A Semi-Automatic Pipeline to Produce a Critical Digital Edition of Literary Texts with Multiple Witnesses out of Text Created through Handwritten Text Recognition. In *Digital Humanities 2022. Responding to Asian Diversity* (pp. 690–691). Tokyo: DH2022 Local Organizing Committee. URL: https://dh2022.dhii.asia/dh2022bookofabsts.pdf (Accessed: 15 June 2024).

Stökl Ben Ezra, D., Rustow, M., & Witty, D. (2022b). Segmentation Mode for Archival Documents with Highly Complex Layout. In *Conference Documents anciens et reconnaissance automatique des écritures manuscrites.* École national des chartes, Paris: YouTube. URL: https://www.youtube.com/watch?v=dE1XUXiuitU (7:07–7:30). (Accessed: 15 June 2024).

Ströbel, Ph., Hodel, T., Fischer, A., Scius, A., Wolf, B., Janka, A., Widmer, J., Scheurer, P., & Volk, M. (2023). Bullingers Briefwechsel zugänglich machen. Stand der Handschriftenerkennung. In A. Busch, & P. Trilcke (Eds.), *DHd 2023. Open Humanities, Open Culture* (pp. 98–102). Belval/Trier: Zenodo. DOI: https://doi.org/10.5281/zenodo.7688631 (Accessed: 15 June 2024).

Wolf, L., Littman, R., Mayer, N., German, T., Dershowitz, N., Shweka, R., & Choueka, Y. (2010). Identifying Join Candidates in the Cairo Genizah, *International Journal of Computer Vision,* 94(1), 118–135.

Literature Related

Camps, J.-B., Vidal-Gorène, Ch., & Vernet, M. (2021). Handling Heavily Abbreviated Manuscripts. HTR Engines vs Text Normalisation Approaches. In E.H. Barney Smith & U. Pal (Eds.), *Document Analysis and Recognition – ICDAR 2021 Workshops. ICDAR 2021* (pp. 306–316). Cham: Springer [= *Lecture Notes in Computer Science,* 12917]. DOI: https://doi.org/10.1007/978-3-030-86159-9_21 (Accessed: 15 June 2024).

Chagué, A., & Thibault, C. (2023). I'm here to fight for ground truth. HTR-United, a solution towards a common for HTR training data. In *Digital Humanities*

2023. Collaboration as Opportunity. Graz: Zenodo. DOI: https://doi.org/10.5281/zenodo.8107449 (Accessed: 15 June 2024).

Clérice, Th. (2022). You Actually Look Twice At it (YALTAi). Using an object detection approach instead of region segmentation within the Kraken engine, *Journal of Data Mining and Digital Humanities,* 1–13. DOI: https://doi.org/10.48550/arXiv.2207.11230 (Accessed: 15 June 2024).

Perdiki, E. (2023). List of manuscripts containing John Chrysostom's Homilies and the relevant manual transcriptions, 1(2). *Zenodo.* DOI: https://doi.org/10.5281/zenodo.8102662 (Accessed: 15 June 2024).

Popović, M., Dhali, M.A., & Schomaker, L. (2023). Artificial intelligence based writer identification generates new evidence for the unknown scribes of the Dead Sea Scrolls exemplified by the Great Isaiah Scroll (1QIsaa), *PLoS ONE,* 16(4. e0249769), 1–28. DOI: https://doi.org/10.1371/journal.pone.0249769 (Accessed: 15 June 2024).

Vidal-Gorène, Ch., Dupin B., Decours-Perez A., & Riccioli T. (2021a). A Modular and Automated Annotation Platform for Handwritings. Evaluation on Under-Resourced Languages. In Lladós, J., Lopresti, D., Seiichi, U. (Eds.), *Document Analysis and Rcognition. ICDAR 2021* (pp. 507–522). Lausanne: Springer. DOI: https://doi.org/10.1007/978-3-030-86334-0_33 (Accessed: 15 June 2024).

Vidal-Gorène, Ch., & Decours-Perez, A. (2021b) A Computational Approach of Armenian Paleography. In E.H. Barney Smith & U. Pal (Eds.), *Document Analysis and Recognition – ICDAR 2021 Workshops. ICDAR 2021* (pp. 295–305). Cham: Springer [= *Lecture Notes in Computer Science,* 12917]. DOI: https://doi.org/10.1007/978-3-030-86159-9_20 (Accessed: 15 June 2024).

Wick, Ch., Reul, Ch., & Puppe, F. (2018). Comparison of OCR Accuracy on Early Printed Books using the Open Source Engines Calamari and OCRopus, *Journal of Language Technology and Computational Linguistics,* 33(1), 79–96.

Figure Credits

Fig. 1: Abraham Lincoln as a grey scale image. CC0. The image was probably originally created by Leon Harmon in 1971, who wanted to find out how much visual information an image could do without in order to still be recognisable.

Fig. 2: Wikipedia © "BrunelloN" CC-BY-SA 4.0

Fig. 3: Screenshot Daniel Stökl Ben Ezra. Manuscript © CC-BY Cambridge University Library

Funded by the European Union (ERC, MiDRASH, Project No. 101071829). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Image Digitization

Hubertus Kohle

https://orcid.org/0000-0003-3162-1304

Abstract A picture is worth a thousand words, as the saying goes. And yet – or precisely because of this – it has been marginalized in European intellectual history or branded as misleading and unreliable. In the Digital Humanities, too, there has been a delay in devoting attention to the image, which is itself reflected in its constitution at German universities. The article deals with the technical characteristics of the digital image, its fundamental peculiarities as an entity of pixel sets with different resolutions and raster graphics as opposed to vector graphics. This is followed by considerations on the structure of image databases and their elementary properties as well as the description of complex search options across astronomically large amounts of image data, which are also becoming increasingly semantically accurate as part of the development of multimodal large language/image/audio models. The presentation concludes with some cursory considerations on AI-controlled image generators.*

Keywords Digital Image, Image Databases, Digital Image Analysis

1. Introduction

The prioritization of the word in European intellectual history and, to a certain extent, the consistently prevailing skepticism towards the image, among others also in the field of reformed theologies, are likewise common in the history of the Digital Humanities. This field of study began in the middle of the 20th century in a philosophical-theological context with the work of the Jesuit Roberto Busa (Father Busa), who used his relationship with Thomas J. Watson, the founder, and head of the *International Business Corporation* (IBM), to use the computer to create an index of the works of Thomas Aquinas. Obviously, the computer was used exclusively for text processing. Not until a quarter of a century later, did the image begin to be used, initially more with textual metadata. In the 1980s, the image as such received research attention (Vaughan 1987, 215–221).

The reasons for this development are both ideological and technical. In the European context and the European history of philosophy, which, according to Alfred

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

Hubertus Kohle: Image Digitization, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 117–126. DOI: https://doi.org/10.11588/heibooks.1521.c21943

North Whitehead, is a series of "footnotes to Plato," the word and its proximity to the idea take precedence. However, the fact that an image is technically more complex to process digitally is more relevant. Vaughan once said he was unable to test his image-oriented search system (called "Morelli") because there were hardly any large quantities of digital images available at the time of his invention.[1] There were reasons for the lack of availability: A reasonably high-resolution image – the meaning of which will be shown – could be several hundred KB in size, and were the image close to slide film quality, the image could easily be several MB. High-resolution images, such as those in certain cases by Google's Arts Project,[2] even reached several hundred MB. Storage space was scarce in the 1980s, and a 20 MB hard drive costed several thousand marks. Yet, a large hard drive could easily hold 20 books with only written text.

Despite technological advances, the positionality of the image itself within the Digital Humanities remains marginal. These are mostly located in linguistics and literature and are therefore purely text oriented. Only in a few cases do academic fields investigate the field of objects, yet do not primarily get into the field of images. Recently, there have been more professorships in the field of visual computing being advertised within the Digital Humanities, despite their popularity in the general field of Computer Science.[3] The focus is on augmented and virtual reality, machine vision, and image processing. Applications can be found primarily in medicine, architecture, design, and gaming. In most cases, AI methods are crucial for successful applications.

2. Technical Basics

Since the late 19th century, photographic reproductions based on originals have determined teaching and research practice, such as in art history, which is the most important auxiliary academic field for theology in the field of images (cf. Dilly 1979). Prior to the 19th century, there were graphic techniques used on a large scale for the reproduction of works of art, transforming the exclusive unique specimen into one of social communication. Copperplate engraving was used in the first place for these types of image reproductions. In the late 20th century, digital reproductions became an additional technique used, which quickly replaced traditional analog forms of production. Currently, chemical-based slide libraries are disposed of, moved to university archives, or stored in the basement. That they are sometimes the subject of art-historiographical and scientific analyses proves rather than disproves their marginality.

1 Verbal communication.
2 See https://artsandculture.google.com (Accessed: 16 June 2024).
3 The DFG has been funding an interdisciplinary priority program with a focus on cultural studies that is dedicated to digital imaging for several years (https://www.digitalesbild.gwi.uni-muenchen.de [Accessed: 16 June 2024]).

The digital image differs from the analog image in that the digital image is a discrete collection of pixels (picture elements) arranged in vertical and horizontal rows in a grid, which are created in an electronic exposure process (see Besser et al. 2003). To say a digital image is a “discrete” collection means that the transition from one pixel to another is made in jumps, whereas it is continuous in an analog image. With digital cameras and scanners, devices have been available since the 1980s that carry out this exposure process either synchronously or diachronically. Scanners are available for a range of different analog source products, such as slide scanners used to digitize analog slides.

Related to the fact that the computer only deals with zeros and ones (this only in parenthesis), the discrete organization of the pixels is often understood from the humanities as an indication that digitally supported analyses only research unambiguity and leave no room for ambiguous results. However, if one considers that the discrete approaches the analog asymptotically as the resolution increases, then there is only a theoretical difference. The related problem of blurring has become a topic of discussion in the Digital Humanities (see Borek et al. 2022).

Depending on the resolution, the raster consists of a different number of pixels, whereby the greater the number of pixels, the greater the detail of an image. One unit of measurement for resolution is dpi (dots per inch), which denotes the number of pixels per inch = 2.54 cm. The size of the original is a decisive factor. If a 2.54 × 1.27 meter oil painting is digitized at 100 dpi, then there are 50 million dots (100 × 100 for the height × 50 × 100 for the width); if I do the same with an analog reproduction of the painting measuring only 25.4 × 12.7 cm, one gets 500,000 dots (10 × 100 for the height × 5 × 100 for the width). In addition to black-and-white and grey-scale images, in which the former is reduced to either black or white, and different brightness values are generated in the latter, color images are currently prevalent, which are intended to produce the most realistic representation of the world. The greater the color depth, the more varied the colors are displayed. The predominant color model in which colors are combined is the RGB color model, in which the primary colors red, blue, and green can be mixed additively in all color tones, i. e., by superimposing them. Realistic color effects are achieved when each of the three colors is coded with 8 bits = 256 colors, which results in a total of 16.7 million colors. High color depth (256 × 256 × 256) combined with high resolution produces large files. While storing and distributing these files is done using modern hard disks with large storage capacities and the large network bandwidth, the calculation of these images is still a task requiring the most powerful machines. The task is simplified though, given that the images can be saved in compressed form, which can reduce their size by a two-digit factor, but the common jp(e)g format *(Joint Photographic Experts Group)* also leads to a loss of quality initially barely visible to the human eye yet can be noticeable with greater compression. The compression level can be preselected when processing the files.

In furtherance of the above-mentioned jpg format, several other formats have been developed, starting with the unchanged original formats. These include bmp *(bitmap)* and tiff *(tagged image file format);* a tiff file can also be compressed. In contrast to jpg, png *(portable network graphics)* does not lose quality upon compression, although the file size is also compressed – albeit to a lesser extent. The gif *(graphics interchange format)* is suitable for moving animations. The manufacturer specific *raw* format contains the most information, which is preferred by professionals as an output format for image processing, but like bmp requires a lot of storage space.

In addition to raster graphics, there are vector graphics, which do not encode each individual point, but mathematically define the geometric and color nature of a pictorial object. The latter has advantages and disadvantages: if a uniform red line does not have to be represented by a whole number of equally colored (red) pixels, but as an instruction, making all points between coordinate x and coordinate y equally red, then it requires less memory space and is more scalable, i.e., the image can be enlarged without loss. With a raster graphic, the appearance in this case becomes less clear because the individual pixels emerge more and more clearly. The disadvantage of vector graphics is a slower processing time, since a coordinate is not visible as a number and must first be converted (rendered) back into a pixel. Incidentally, the principle of vector graphics works best with schematic visualizations with few local color variations.

3. Image Databases

Within the Visual Arts, there is a series of databases in which digital reproductions of two- and three-dimensional works of art are stored. Museum directories are a primary storage location, with the advantage that the digital reproductions are most likely produced according to the originals and not according to printed reproductions, where moiré effects repeatedly occur, one of the negative effects of digitizing a paper print. Comprehensive databases are available primarily in the American *artstor*[4] and in the *Foto Marburg* database[5], which also has historical relevance, as it ultimately draws on photo campaigns dating back to the early 20th century and therefore partly includes objects that have not yet been affected by the potentially extensive losses of the two World Wars. There is also *Prometheus*[6], which represents the synthesis of over a hundred individual directories searchable by a common interface. All large collective databases, not referring to individual museums but to entire clusters of museums or architecture, contain several million images and are continually expanding their holdings.

4 See https://www.artstor.org (Accessed: 16 June 2024).
5 See https://www.bildindex.de (Accessed: 16 June 2024).
6 See https://www.prometheus-bildarchiv.de (Accessed: 16 June 2024).

The *Prometheus* database is predominant in the German-speaking world, partly because it is inexpensive compared to the American databases. The principle of this meta-database, which does not provide any data of its own apart from the search functionality, is suitable for the peculiarities of the internet. It is based on a notion of “shared authorship” and relies on the influence of the crowd, which in total is unbeatable. However, this principle is not without disadvantages, as the quality of the data provided by the individual suppliers is rarely checked to reach a critical mass as quickly as possible. The makers of *Prometheus* recognized this limitation at the outset, just like all the internet entrepreneurs who invest heavily in the rapid generation of large numbers of users: if one wishes to penetrate the market – which in the case of *Prometheus* means, above all, being used by university institutions – one must make a number of artwork reproductions that can be used primarily for teaching purposes available very quickly. Only then will teachers feel compelled and justified to convert the projection system from slide to digital – which has now occurred on a large scale. As mentioned, doing so comes at a cost, namely, the at times poor quality of the images and the doubling or tripling four to fivefold increase in reproductions after better-known works of art. The creators of *Prometheus* have responded to the issue by asking users to mark the individual images with asterisks so that the poorer ones can be eliminated, a procedure ingeniously relying again on “collective wisdom” (style-defining: Surowiecki 2005). Reducing the number of poor-quality images could also be accomplished using the procedure based on SIFT proposed by Schneider (2019). That *Prometheus* increasingly integrates the databases of individual large museums likewise serves to increase the quality.

The resolvable disadvantages are offset by a significant advantage resulting from the distribution of the work over several parties, which in turn can be used everywhere on the network. In the past, any institution involved with images had its own archive. The limited workforce led to a patchy stock of images, noticeable even in extensive collections. If hundreds (if not thousands) of digitizers are active across the network, then there are much larger amounts of data/images able to be used by any (registered) internet user, despite any duplications. The joint use of the various, often technically divergent, systems has been successfully realized for some years with software solutions such as IIIF *(International Image Interoperability Framework).*[7]

The data is entered into a relational database where the information is organized in tables that are related to each other and queried using sql *(search query language).* Recently, graph databases are coming into the limelight provided their flexibility. In them, data entities called *knots* (e.g., objects, but also names, etc.) are defined, which are related to each other via *edges* and stored in rdf *(resource description framework)* triples. The system can reach any level of complexity.

7 See https://iiif.io (Accessed: 16 June 2024).

4. Image Search

One of the advantages of digitization in the field of image databases is that individual objects can not only be located in a specific place but can also be accessed in a differentiated way depending on the level of detail of the description data. In the traditional slide library, Michelangelo's "Moses" is found under the artist's name, which is of little help to anyone interested in the iconography of Moses, if one does not know that Michelangelo also devoted himself to the subject. Of course, it is possible to set up an alternative category "Iconography" and classify "Moses," but someone interested in Michelangelo who does not know that he created a "Moses" might not find what they are looking for. The solution is to place two reproductions of the same work under both the artist and the iconography. However, doing so would inflate the picture collection, and if a category of date of creation were added (in this case approx. 1514) under which all works from this period were categorized, or one of the preserving institutions (here the church of San Pietro in Vincoli in Rome), then the number could be increased three or fourfold. From a cost and space perspective, this solution would not be desirable.

The digital realm is different. Each work can be digitally indexed in a variety of different ways without needing to be stored multiple times. If the work is registered in a database in which, artist, title, date, technique, and storage location are stored, then each of these categories can be searched, even in combination. A search for the date "1514," for example, would return all the works stored in this database from the year 1514. Then, when combined with the artist "Michelangelo," everything that the Renaissance artist created in that year will be shown. In this way, statistically oriented investigations are possible such as: which museum/church (repository) owns the most works by Michelangelo? In which period did Michelangelo create the most works? What are Michelangelo's preferred themes? The latter, however, is somewhat less trivial: Work titles are ambiguous and thus cannot be determined by simply naming the title; it can be "Maria," "Mother of God," or even "Mary." A classification system such as *iconclass*[8] can eliminate this ambiguity and the different names can be *mapped* onto each other. A relatively easy-to-use tool for carrying out demanding tasks is available in *Museum Analytics*[9], which was programmed expressly for interested parties who do not come from IT. It remains to be seen whether LLMs *(Large Language Models),* which are currently in vogue in the form of ChatGPT, e.g., will be able to simplify the research possibilities.

The high number of reproductions in these image databases in conjunction with their content indexing can lead to a methodological decanonization. While under analog conditions the access possibilities were limited and depended largely on the knowledge of the searcher, so that when visiting the card catalog, one needed to know

8 See https://iconclass.org (Accessed: 16 June 2024).

9 See https://dhvlab.gwi.uni-muenchen.de/max (Accessed: 16 June 2024).

in advance the intended outcome, the digital age has changed the process entirely. If the media logic of digitization is followed and it is not seen as a simple extension of analogization, there are more diverse search options, implying that the search results are also diverse. Even if the old rule that new media imitate the old applies on both the supply and, above all, on the demand side, so that the methodical transformation remains manageable, then the tendency towards diversification will increase as the range of offerings continues to differentiate.

5. Image Calculations

From a purely practical perspective, digitization offers considerable advantages in the management of and search for images. However, image digitization also enables new analytical possibilities, which undoubtedly represent a paradigm shift in the field. Until now, it has been possible to describe images, identify their forms and content, and place them in a historical or stylistic context. However, it was not possible to go into them directly, or even indirectly. A digital image is not just a grid of pixels; each individual pixel can be principally addressed, identified, and changed. Whereas the search options described above were limited to the information referred to as metadata, i.e. information added by a human or technical intelligence, here there is the completely new possibility of focusing on the object itself, not the secondary information about it, e.g. in a histogram, with which the distribution of color and brightness values in an image can be displayed diagrammatically. The conclusions are intriguing.

The computer is a formalistic machine; it recognizes color dots at a particular location, and nothing more, but it does so with tremendous speed and accuracy yet without recognizing semantics. For the computer, the image of a dog is simply a cloud of pixels. In Art History, formalism has once again received attention, which would progress further by the image-addressing procedures in the digital age. Heinrich Wölfflin, who achieved worldwide success more than a hundred years ago with his *Kunstgeschichtliche Grundbegriffe [Basic Concepts of Art History]* and distinguished works of art from the Renaissance and Baroque periods purely on the basis of their formal characteristics, is experiencing a renaissance in this context (Wasielewski 2023). Media scholars assume a freeing of the image from the clutches of human categorizations, which ultimately leads to a de-historicization, as historical thinking is ultimately founded in human language (Pias 2003). Along with this intimate connection, the problem takes on an almost eschatological meaning, often associated with the discourses of Posthumanism.

Just as in *Optical Character Recognition* (OCR) of texts, where the available letters as pixel clouds are given meaning in a training process, these pixel clouds can also be trained for dogs and all other objects, which is currently being done using neural networks trained for semantics in a machine learning process. These neural networks are modeled on neuronal connections in the brain and, in the eyes of techno-utopians, are still less powerful than the brain, mainly because the latter still enables far more nerve cell connections than even the most powerful computers. By way of example, one of the consequences is that a small child only needs to be shown a few images of a dog to gain an abstract idea of the dog and to recognize dogs that do not correspond exactly to the examples shown, whereas the computer must be given several hundred, if not thousands, of these examples to derive a general model.

For image searches, this means that going forward the searches will increasingly be based on the images themselves rather than on metadata. From a purely technical point of view, this means that larger quantities of images can also be searched, i.e., those that go beyond art databases containing millions of images. The problem sounds purely theoretical, but when one considers that 350 million photos are uploaded to Facebook alone every day, the practical side quickly becomes apparent. However, much research needs to be done in Art History in order to present a functional and reputable search system based on direct image addressing. It is already possible to recognize a person in a picture, even a person pouring water over another person's head is no longer an insurmountable problem for advanced AI programs. But recognizing John the Baptist baptizing Jesus Christ in the Jordan River is hardly feasible with the existing approaches. To do so, one needs to increase massively the amount of training data and supplement the 400 million image-text combinations used to train models such as CLIP from the Open AI Initiative, especially with more specific combinations. In the case of individual sciences, this result could primarily be achieved by drawing on a large amount of expert knowledge available in the form of machine-readable research texts. Such approaches also show how important it is to publish scientific texts as comprehensively as possible in open access regarding potential solutions.

6. Image Forgeries

AI cannot only be used directly to address images. It can also be used to produce images that no longer have a depictive reference to reality. This was known before as well, but only as image modifications, not as completely new inventions. These depictions are thus referred to as fake images or deep fakes. AI supported systems such as *DALL-E, Midjourney,* or *stable diffusion,* in addition to serious applications in which the non-realistic focus cannot be doubted, allow such uses. The sociopolitical effects remain to be seen. If a picture of the Pope running away from the police recently went viral on the internet, then a well-informed individual should recognize the falsity of the image, just because of its extreme improbability. However, a picture of the same Pope amused by a caricature of Mohammed is likely to cause displeasure in the Islamic world, even if the Pope has never actually seen such caricatures.

Literature Cited

Besser, H. (2003). *Introduction to Imaging. Revised Edition.* Ed. by S. Hubbard & D. Lenert. Los Angeles: Getty Research Institute. URL: https://www.getty.edu/publications/virtuallibrary/0892367334.html (Accessed: 16 June 2024).

Borek, L., Zeppezauer-Wachauer, K., & Ketschik, N. (2022). Eindeutig Uneindeutig. Zur Modellierung von Unschärfe in der Mediävistik. In *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte* [Blog]. DOI: https://doi.org/10.58079/rha5 (Accessed: 16 June 2024).

Dilly, H. (1979). *Kunstgeschichte als Institution. Studien zur Geschichte einer Disziplin.* Frankfurt a. M.: Suhrkamp.

Pias, C. (2003). Das digitale Bild gibt es nicht. Über das (Nicht-)Wissen der Bilder und die informatische Illusion, *zeitenblicke,* 2(1), no pag. URL: https://www.zeitenblicke.de/2003/01/pias/pias.pdf (Accessed: 16 June 2024).

Schneider, S. (2019). Über die Ungleichheit im Gleichen. Erkennung unterschiedlicher Reproduktionen desselben Objekts in kunsthistorischen Bildbeständen. In *DHd 2019. Digital Humanities multimedial und multimodal.* Frankfurt a. M./Mainz: Zenodo. DOI: https://doi.org/10.5281/zenodo.4622051 (Accessed: 16 June 2024).

Surowiecki, J. (2004). *The Wisdom of Crowds. Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economics, Society and Nations.* New York: Doubleday.

Vaughan, W. (1987). The Automated Connoisseur. Image Analysis and Art History. In P. Denley, & D. Hopkin (Eds.), *History and Computing* (pp. 215–221). Manchester: Manchester University Press.

Wasielewski, A. (2023). *Computational Formalism. Art History and Machine Learning.* Cambridge, Mass.: MIT Press.

Wölfflin, H. (1915). *Kunstgeschichtliche Grundbegriffe. Das Problem der Stilentwickelung in der neueren Kunst.* Munich: Bruckmann. DOI: https://doi.org/10.11588/diglit.27250 (Accessed: 16 June 2024).

Literature Related

Altenhöner, R., Berger, A., Bracht, Ch., Klimpel, P., Meyer, S., Neuburger, A., Stäcker, Th., & Stein, R. (2023). DFG-Praxisregeln “Digitalisierung”. Aktualisierte Fassung 2022. Online: *Zenodo.* DOI: https://doi.org/10.5281/zenodo.7435724 (Accessed: 16 June 2024).

Brown, K. (2020). *The Routledge Companion to Digital Humanities and Art History,* New York/London: Routledge.

Cohen, D., & Rosenzweig, R. (2006). *Digital History. A Guide to Gathering, Preserving, and Presenting the Past on the Web.* Philadelphia: University of Pennsylvania Press.
Jannidis, F., Kohle, H., & Rehbein, M. (Eds.). (2017). *Digital Humanities. Eine Einführung.* Stuttgart: J.B. Metzler Verlag.
Kuroczynski, P., Bell, P., Dieckmann, L., Hoppe, S., & Münster, S. (Eds.). (2018), *Computing Art Reader. Einführung in die digitale Kunstgeschichte.* Heidelberg: arthistoricum.net [= *Computing in art and architecture,* 1].

Computational Audio and Music Analysis

Christof Weiß

https://orcid.org/0000-0003-2143-4679

Abstract With the ongoing digitization, not only textual documents but also other types of media have become available in large quantities. This includes audio recordings comprising three main types of content: speech, environmental sounds (e.g. natural or urban soundscapes), and music. While all may be relevant for theological research, this chapter focuses on using audio recordings for studying sacred music (Computational Musicology). After introducing fundamentals of audio data, we first outline a technique for visualizing the tonal content (local keys and modulations) within a music recording and apply this technique to Bach's *Johannespassion* BWV 245. Second, we demonstrate the potential of audio recordings for corpus analysis. We present an approach for studying the tonal complexity and its evolution over centuries. With this technique, we examine the tonal evolution of sacred music exploiting an annotated audio corpus (5,773 tracks) stemming from a leading music publisher for choral music, the Carus-Verlag Stuttgart.

Keywords Audio Signal Processing, Harmony Analysis, Computational Musicology, Corpus Analysis

1. Audio Data and Applications

Ongoing digitization efforts result in an increasing number of archives and corpora on cultural artifacts. Textual data have been the starting point for the Computational Humanities (CH) by using statistical methods on comprehensive literary texts. Nowadays, further modalities are available in the same vein, including audio (sound) recordings. In contrast to text, raw audio poses a number of challenges: First, due to a considerably larger size (one second of uncompressed stereo audio corresponds to 88,200 16-bit values), audio storage and transmission demands for more resources – a problem that has been addressed by efficient audio coding technology beginning with MP3 audio compression and similar codecs. Second, the computational analysis of audio data requires more elaborate processing techniques. In contrast to text, explicit symbols such as characters or words (in speech) or note events (in music) are not directly accessible from audio. To extract this information, algorithmic solutions have been developed for decades, comprising techniques from engineering (signal processing) and computer science (pattern recognition, machine learning, and nowadays deep learning/AI technology). Central venues for this research are the *International*

Christof Weiß: Computational Audio and Music Analysis, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 127–145. DOI: https://doi.org/10.11588/heibooks.1521.c21944

Conference on Acoustics, Speech, and Signal Processing (ICASSP)[1] or the IEEE/ACM *Transactions on Audio, Speech, and Language Processing.*[2]

In general, audio data contains a mixture of various different sources. Consider, for example, the case of a movie sound track or an audio book, which may comprise speech (by different speakers), background music, as well as diegetic sounds (sound effects, sounds generated by people or objects in the plot, or music played or sung within the plot). The separation of these sources is a major computational challenge (Smaragdis 2004). In many cases, however, we are dealing with clean audio data, covering one of three types of content (speech, environmental sound, music), which we shortly summarize in the following.

Speech data. Since the most fundamental type of audio is spoken language, a large part of audio technology is motivated by applications for inter-personal communication. Consequently, speech processing has driven the development of digital audio technology with research on fundamental time-frequency transforms, specific audio features such as *Mel-frequency cepstral coefficients* (MFCCs), dynamic programming techniques such as *Hidden Markov Models*, and more recently, machine-learning algorithms based on neural networks (Bäckström et al. 2022). A central venue for this research field is the annual INTERSPEECH conference by the *International Speech Communication Association* (ICSA).[3] Motivated by different applications, speech processing comprises a variety of tasks such as speech coding and transmission, speaker identification, speech-to-text transcription, analysis of emotion, prosody, or dialect, or audio forensics. More recently, the generation of coherent speech signals from text or directly from a user query has matured due to tremendous progress in generative deep-learning techniques. Interactive voice assistants are one of the most prominent applications of such technology. For CH, efficient speech-to-text (or automatic speech recognition, ASR) systems (Schneider et al. 2019) are of high interest since they can be used as a preprocessing step for the subsequent application of text-based CH strategies.

Environmental sounds. Besides speech, a second field of study covers the processing of sounds in a more general sense. There is a dedicated research community on the detection and classification of acoustic scenes and events (DCASE),[4] which addresses a variety of sound event detection and acoustic scene classification tasks within the annual DCASE challenge. One prominent application of such technology is wildlife and biodiversity monitoring where, for instance, natural reserves are equipped with microphone devices to capture animal sounds for analyzing the presence of species, e.g.,

1 See https://ieeexplore.ieee.org/xpl/conhome/1000002/all-proceedings (Accessed: 21 June 2024).
2 See https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6570655 (Accessed: 21 June 2024).
3 See https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6570655 (Accessed: 21 June 2024).
4 See https://dcase.community (Accessed: 21 June 2024).

of birds (Bardeli et al. 2010). Another application is monitoring urban sound scenes for targeting issues such as noise pollution or detecting crime or potential dangers. This has been done, for instance, in a large project in New York City (Bello et al. 2019). Specific to such applications is the demand for low-resource technology since many signals are recorded in parallel over large areas and very long time spans. Thus, (pre-) processing has to be done locally on individual sensor units (edge devices).

Music audio recordings. The third major category is music audio. In general, music data exists in a variety of digital data types including (beyond audio) graphical sheet music or symbolic (i.e., machine-readable) scores, which explicitly encode musical symbols and usually allow for the most detailed analyses (Temperley 1997; Bellmann 2012; White 2013; Nakamura & Kaneko 2019). However, scores are not available for a variety of music traditions and styles including improvised (e.g. organ improvisation in church), electronically generated, or orally transmitted music. Moreover, audio-based approaches allow for studying performance aspects such as the behavior of congregational singing in church. Finally, symbolic scores are hard to acquire since manual creation is time-consuming and automated conversion from sheet music images (*optical music recognition,* OMR, see Calvo-Zaragoza et al. 2020) or audio recordings (*automatic music transcription,* AMT, see Benetos et al. 2019) to symbolic scores remains often unsatisfactory and demands for considerable manual post-processing. For this reason, audio recordings are a promising alternative since they allow for efficiently scaling up computational music analyses to large corpora (Scherbaum et al. 2017; Mauch et al. 2015; Weiß et al. 2018; 2019). This requires advanced computational techniques that convert the data into semantically meaningful representations that can be directly interpreted by music experts. Such technology is developed within an interdisciplinary research community centered around the *International Society for Music Information Retrieval* (ISMIR),[5] which offers an annual conference and a journal.[6] *Music Information Retrieval* (MIR) comprises a variety of tasks and applications including music synchronization, harmony analysis (chord and key detection), beat and tempo tracking, genre and style classification, audio decomposition, and music transcription (Müller 2021). Beyond these audio analysis tasks, music generation tasks and other musical data types play an important role within MIR.

In the following, we focus on the potential of MIR technology for musicological research, which demands for specific datasets and analysis techniques. We present two analytical studies. The first one (Section 2) deals with the visualization of harmony (local keys and scales) to analyze the tonal organization of large-scale musical works considering Johann Sebastian Bach's *Johannespassion* BWV 245 as an example. The second one (Section 3) demonstrates an audio-based corpus analysis of musical style

5 See https://www.ismir.net (Accessed: 21 June 2024).
6 See https://transactions.ismir.net (Accessed: 21 June 2024).

in Western sacred music relying on a dataset by a leading publisher for choral and sacred music, the Carus-Verlag Stuttgart.

2. Visualizing Tonal Structures: A Case Study on Bach's *Johannespassion*

This section presents an algorithmic approach for visualizing tonal information over the course of an audio recording. Following (Weiß & Müller 2021), we introduce basic notions of audio, fundamental processing techniques, and our visualization strategy at the example of the choral No. 22 "Durch Dein Gefängnis" from J.S. Bach's *Johannespassion* BWV 245. We finally apply this technique to the complete *Johannespassion* and show its potential for studying the tonal organization of large-scale works.

2.1 Extracting Spectral Information from Audio

The starting point of audio analysis is an acoustic waveform (also referred to as *signal*), as shown in Fig. 1a for a recording of the Bach choral. In a first step, we perform a spectral analysis (Müller 2021, Chap. 2). For this purpose, we first divide the signal into local time windows *(frames).* The width of the time window (given in seconds) is a critical parameter that has to be adapted to the particular application requirements since there is a trade-off between frequency and time resolution. Within a frame, the salience of different frequencies is calculated, which can be realized, for instance, by the Fourier transform.[7] This time window is now shifted over the signal, so that one receives for each frame a local frequency distribution. This results in a time-frequency representation, a so-called *spectrogram*, which is shown in Fig. 1b for the Bach choral example.

For tonal analysis, we further summarize this spectral information according to musical pitches. To this end, we make the simplifying assumption that the pitch content can be described well enough by the twelve-tone equal-tempered scale. We further assume that pitch-class information (ignoring a pitch's octave) is sufficient for our tonal analysis tasks. Thus, we end up with the twelve chromatic pitch classes c, c♯, d, d♯, …, b. Here, enharmonic differentiation of pitches such as c♯ and d♭ is not possible. For each frame of the spectrogram, the frequency components are aggregated according to these twelve pitch classes. This results in a time-chroma representation

7 Together with the windowing procedure described before, this specific variant is denoted as (discrete) *short-time Fourier transform* (STFT). Other transforms have been developed for specific applications such as the *constant-Q transform* (CQT) for pitch analysis or the *modified discrete cosine transform* (MDCT) for audio coding.

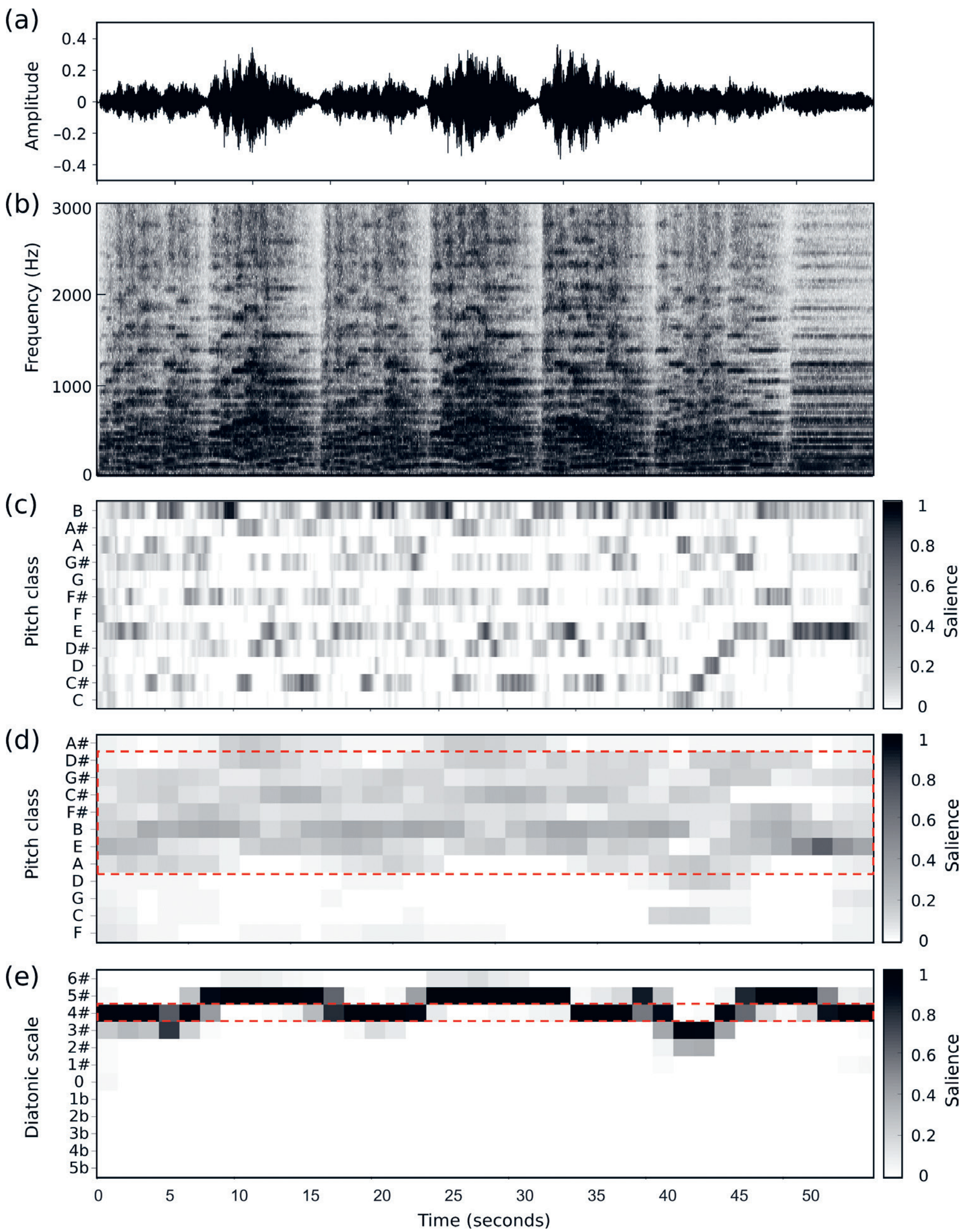


Fig. 1 Choral No. 22 "Durch Dein Gefängnis" from J.S. Bach's *Johannespassion* BWV 245, recording by *The Scholars Baroque Ensemble* (Naxos 1994). (a) Waveform of the audio signal. (b) Spectrogram. (c) Chromagram. (d) Smoothed chromagram, re-ordered according to perfect fifths. (e) Visualization of diatonic scale probabilities.

or chromagram. Fig. 1c shows such a chromagram (time resolution 10 Hz) for the Bach choral example. This representation captures the energy distribution of the audio signal over the twelve chromatic pitch classes over time. Converting music recordings into chromagrams as an intermediate step in the processing chain is a fundamental approach for various MIR tasks such as key estimation, scale analysis, and chord recognition. For details of the underlying signal processing, we refer to Müller (2021).

2.2 Visualizing Diatonic Scales

The observation and measurement of energy distributions in pitch classes provides only limited information for tonally complex polyphonic music. More useful categories for this purpose are intervals, chords, or scales, which have to be described by further processing steps. In the following, we consider the measurement of pitch content according to the twelve diatonic scales. For this purpose, we first smooth the chromagram (by local averaging) in order to account for the coarser time resolution of musical scales (the pitches of a scale do usually not occur within a time span as short as our 100 ms chromagram frames). This is indicated by Fig. 1d, where we chose a filter length of 45 frames (i.e., 4.5 seconds). Second, a chromagram for each of these smoothed frames is compared with binary templates corresponding to the scales. For example, the 0 diatonic (pitch content of C major or A minor, no accidentals) is modeled by a template in which the values for the seven pitch classes c, d, e, f, g, a, h are set to 1 and for the remaining five pitch classes c♯, d♯, f♯, g♯, a♯, to 0. Fig. 1d highlights in red the pitch content of the 4♯ scale (corresponding to E major, the global key of the choral). Due to the similarity of diatonic scales that are related by a perfect fifth (which share six out of seven pitch classes), we organize the scales according to the circle of fifths.

By locally matching the chromagram frames with the twelve different templates, probabilities for the occurrence of these scales over time can be computed. This leads to a generalized time-diatonic representation where the probabilities can be visualized via a grayscale scheme (here, black corresponds to probability 1 and white to probability 0). Fig. 1e shows such a visualization for a recording of the Bach choral. Additional processing steps may be required to better highlight important structures. For example, further temporal smoothing or enhancement of the higher energy values and suppression of low, noise-like values can lead to clearer visual structures. The latter is realized by means of an exponential rescaling of the energy values (similar to the *Softmax* function).

The time-diatonic representation derived from a music recording is initially organized according to physical time steps (seconds). For some applications, such as evaluating the musical form of large-scale works, this can be useful. However, for comparison with score representations or other interpretations of a work (cross-version analysis), it is useful to consider musical time information (e.g., structural boundaries or measures/beats). If such information is available, the temporal components of the time-diatonic representation can be musically smoothed in order to obtain, for example, a representation with quarter note resolution.

2.3 Bach Choral Example

Let us now take a closer look at the visualization result for our Bach choral (Fig. 1e). The first part starts in the choral's global key (4♯) and modulates into the upper fifth key or dominant key (5♯) starting at about 8 seconds. An interesting observation is the deviation from the 4♯ scale at roughly 5 seconds, where alterations (here the d in the chord g♯-d-e-h) affect the result. In fact, no d♯ is found in the entire first measure, so a high probability is visible for the 3♯ diatonic as well. This shows that our procedure does not provide an explicit *recognition* of keys, but only describes the local pitch content in terms of diatonic scales. On the basis of the time-diatonic representation, the rough harmonic progression of the Bach choral can be followed conveniently. The beginning phrase (4♯ with modulation to 5♯) is repeated with other lyrics. The choral proceeds with more complex harmonies. Here, the chromatic passage in the bass (text "Unsere Knechtschaft") at about 40 seconds results in several scales obtaining non-zero probabilities. Finally, the choral ends in the global key of E major (4♯). Beyond the easy access to music recordings and their straightforward processing, the direct applicability to audio is of particular advantage since the analysis can be directly linked to the acoustic impression, e.g., by using a running cursor as animation.

2.4 Large-Scale Tonal Visualization of the *Johannespassion*

One of the major benefits of the presented analysis strategy is its scalability. Large-scale works such as operas, oratorios, or symphonies can be visualized in a compact and consistent way, thus allowing for a good overview of their tonal conception – a very important aspect since tonality has been a central means for formally structuring long works. We now demonstrate this by extending our analysis from the one-minute choral to the whole *Johannespassion* BWV 245, which (in the performance by *Scholars Baroque*) amounts to a total playing time of roughly two hours. Fig. 2 shows the result of this analysis, with the whole work being condensed to a single plot.

To avoid micro-fluctuations obscuring the coarse-scale tonal structure, we now opt for a much larger window size when averaging the chroma features before the template matching. As opposed to the more fine-grained analysis of the isolated choral (Section 2.2) with a filter length of 4.5 seconds, we now chose a filter length of 60 seconds. This leads to a suppression of details but enhances the robustness of the method and helps to emphasize the overall tonal structure. Let us now discuss the results. For better readability, Fig. 2 only contains the numbers of the individual movements.

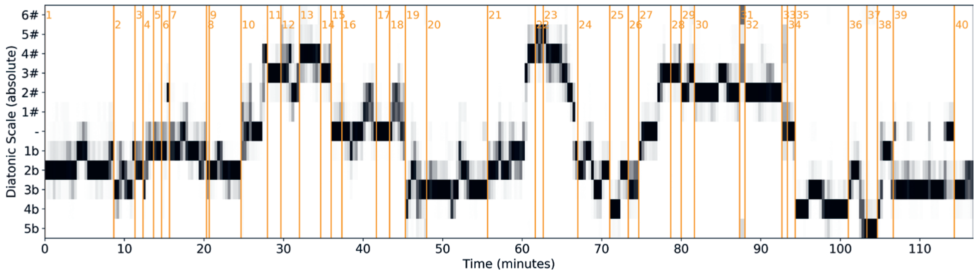


Fig. 2 J.S. Bach's *Johannespassion* BWV 245, complete recording by Scholars Baroque. Diatonic scale visualization for all movements.

The form (instrumentation) and text cues are given in Table 1 for reference. First, we observe a certain local stability. Within a movement and between neighbouring movements, distant modulations are rare, meaning that the next movement usually starts in the same or a closely related key. Second, we identify a certain tonal closure. The work starts in the the −2 diatonic (No. 1, Chorus "Herr, unser Herrscher"in G minor) and ends with two numbers in the −3 diatonic (No. 39, Chorus "Ruht wohl, ihr heiligen Gebeine" in C minor and No. 40, Choral "Ach Herr, laß dein lieb Engelein" in E♭ major.), only a fifth apart. In between, this tonal region is clearly left, with the peak at our choral discussed above, No. 22 "Durch Dein Gefängnis", which is, with its 4♯, about seven fifths apart. Not only is this remarkable since E major has been related to death in the Baroque era – Johann Mattheson ascribed this key a "deadly sadness" ("tödliche Traurigkeit") in his 1713 writing Das *Neu-eröffnete Orchester.* Indeed, the choral is also the center of the passion in several respects. Regarding physical playing time, it occurs right in the middle of the work, after roughly one hour (which is not the beginning of the second part). Moreover, there is a clear symmetry around this choral, which can be nicely observed in our visualization. Numbers 21 and 23 (both a series of recitatives and choruses, which are repeated with different text) include transitions to and from 4♯, respectively. The arias Nos. 20 and 24 each comprise the scales 3♭ and 2♭. Nos. 15–17 and 27 emphasize the 0 diatonic, and so forth. These tonal relationships show the ingenious conception of Bach's *Johannespassion* and his deliberate usage of tonality for structuring the work and for putting emphasis on the theological messages of the passion narrative.

3. Audio-based Corpus Analysis

Scaling analyses to comprehensive works is one major advantage of computational methods. Another highly interesting possibility is to analyze whole corpora of musical works. Here, a corpus may refer to a closed set of works, e.g., all chorals by

Tab. 1 J.S. Bach's *Johannespassion* BWV 245, overview of numbers (movements), forms, and text cues.

No.	Form	Text cue
Parte Prima		
1	Chorus	Herr, unser Herrscher
2	Rezitativ, Chorus	Jesus ging mit seinen Jüngern
3	Choral	O große Lieb
4	Rezitativ	Auf daß das Wort erfüllet würde
5	Choral	Dein Will gescheh
6	Rezitativ	Die Schar aber und der Oberhauptmann
7	Arie	Von den Stricken meiner Sünden
8	Rezitativ	Simon Petrus aber folgete Jesu nach
9	Arie	Ich folge dir gleichfalls
10	Rezitativ	Derselbige Jünger war dem Hohepriester bekannt
11	Choral	Wer hat dich so geschlagen
12	Rezitativ, Chorus	Und Hannas sandte ihn gebunden
13	Arie	Ach, mein Sinn
14	Choral	Petrus, der nicht denkt zurück
Parte Seconda		
15	Choral	Christus, der uns selig macht
16	Rezitativ, Chorus	Da führeten sie Jesum
17	Choral	Ach großer König
18	Rezitativ, Chorus	Da sprach Pilatus zu ihm
19	Arioso	Betrachte, meine Seel
20	Arie	Erwäge, wie sein blutgefärbter Rücken
21	Rezitativ, Chorus	Und die Kriegsknechte flochten eine Krone
22	Choral	Durch dein Gefängnis, Gottes Sohn
23	Rezitativ, Chorus	Die Jüden aber schrieen und sprachen
24	Arie	Eilt, ihr angefochtnen Seelen
25	Rezitativ, Chorus	Allda kreuzigten sie ihn
26	Choral	In meines Herzens Grunde
27	Rezitativ, Chorus	Die Kriegsknechte aber
28	Choral	Er nahm alles wohl in acht
29	Rezitativ	Und von Stund an nahm sie der Jünger
30	Arie	Es ist vollbracht
31	Rezitativ	Und neiget das Haupt
32	Arie	Mein teurer Heiland, laß dich fragen
33	Rezitativ	Und siehe da, der Vorhang im Tempel zerriß
34	Arioso	Mein Herz, in dem die ganze Welt
35	Arie	Zerfließe, mein Herze
36	Rezitativ	Die Jüden aber, dieweil es der Rüsttag war
37	Choral	O hilf, Christe, Gottes Sohn
38	Rezitativ	Darnach bat Pilatum Joseph von Arimathia
39	Chorus	Ruht wohl, ihr heiligen Gebeine
40	Choral	Ach Herr, laß dein lieb Engelein

J.S. Bach or all string quartets by L. van Beethoven. It might, however, also refer to an open subset of a whole time span. To sketch a long-term goal, we envision the analysis of the development of Western sacred music, spanning more than a thousand years from monophonic chant up to today's avantgarde compositions. Following Weiß & Müller (2023), we now present a first step towards such corpus analysis. To this end, we consider an audio dataset provided by a leading publisher of choral and sacred music.

3.1 The Carus Audio Corpus

The Carus-Verlag, founded near Stuttgart, Germany, in 1972 is a family business focusing on vocal and sacred music. Their sheet music editions include around 45,000 works (most of them vocal compositions) and reflect the development of five centuries of choral music, ranging from Gregorian chant, madrigals, and motets of the Renaissance, to contemporary choral music, and works for jazz and pop choir.[8] Carus offers scholarly-critical music editions of the most important oratorios, masses, and cantatas in music history, orientated towards historically informed performance practice. Being also active as a record label, Carus releases reference recordings based on their own editions. The CAC comprises the majority of the Carus CD releases (as of 2019), totalling 7,115 tracks corresponding to individual works or movements (for multi-movement works). Since we want to focus on original art music compositions, we perform a first cleaning step where we remove works without composer, works without composer life dates, arrangements, pop music, children songs, and christmas songs. After this, 5,773 tracks (movements) remain belonging to 2,409 different works with a total duration of 389:52:20 (hh:mm:ss). On average, a work has 2.4 movements and a duration of 9:43 (mm:ss). However, we note that the number of movements per work is highly unbalanced, with many one-movement works on the one hand and many large-scale works (such as the *Johannespassion*) on the other hand. Table 2 provides statistics over the CAC's annotations at the *work level,* where information such as key or instrumentation always refer to the overarching composition. Roughly half of the works has annotations regarding the year of composition (work date). The majority is annotated regarding instrumentation. As said, there is a strong focus on vocal music in general as well as on choral music.[9] From the perspective of tonal analysis, the availability of key annotations for roughly half of the works (1,166 out of 2,409) is of particular relevance. There is a bias towards major keys as well as a considerable number of other keys (church modes). As mentioned above, CAC spans roughly 450 years, covering the period from about 1570–2020. In

8 See https://www.carus-verlag.com/en/ueber-carus (Accessed: 21 June 2024).

9 Please note that, due to the work-related annotations, individual solo vocal movements (e.g., an aria) within a choir work (e.g., an oratorio) are counted towards choral works.

Tab. 2 Statistics of CAC and its annotations. All numbers refer to full works (not individual movements).

Annotation type	No. of works
– All –	2,409
Work date	1,151
Instrumentation	1,964
instrumental	200
vocal	1,764
choral	1,400
solo	364
Key	1,166
major	673
minor	348
other	145

total, the works stem from 234 different composers. Fig. 3 shows a historical view on the composer dates for composers with at least five works. Well-known composers like F. Mendelssohn-Bartholdy, J. S. Bach, or W. A. Mozart contribute a significant part. However, CAC also comprises less known composers such as H. Schütz or M. Reger. Carus even makes great efforts to bring almost forgotten works by G. A. Homilius or J. G. Rheinberger back into the focus of the choir scene. A particular interesting fact is the good coverage of the late 15th and 16th century. In the 20th century, in contrast, we find a lower number of works, almost observing a gap around 1950.

3.2 Work Count Curves and Approximation Strategy

To analyze musical styles in their historical context, one ideally has information about the true *work dates*, i. e., the year where a composition was completed. Musical styles may evolve rapidly, and composing is subject to trends, being influenced by other composers, the taste of audiences, or extra-musical stimuli. One might think of composers with several *creative periods,* such as L. van Beethoven or A. Schönberg. However, collecting reliable work date annotations for larger datasets requires a substantial amount of manual research, and this information is unknown or in doubt for quite a number of works. Even if one knows all composition dates, it becomes difficult to create a dataset with a balanced coverage of all years. Because of such problems, in our previous work (Weiß et al. 2019), we adopted a pragmatic approach by projecting works onto the historical time axis based on *composer dates,* i. e., the information on birth and death year, which is considerably faster to acquire. To approximation of work counts over the course of a composer's life, we assumed that a composer starts composing not before a certain (fixed) age. For the remaining years (ages), we com-

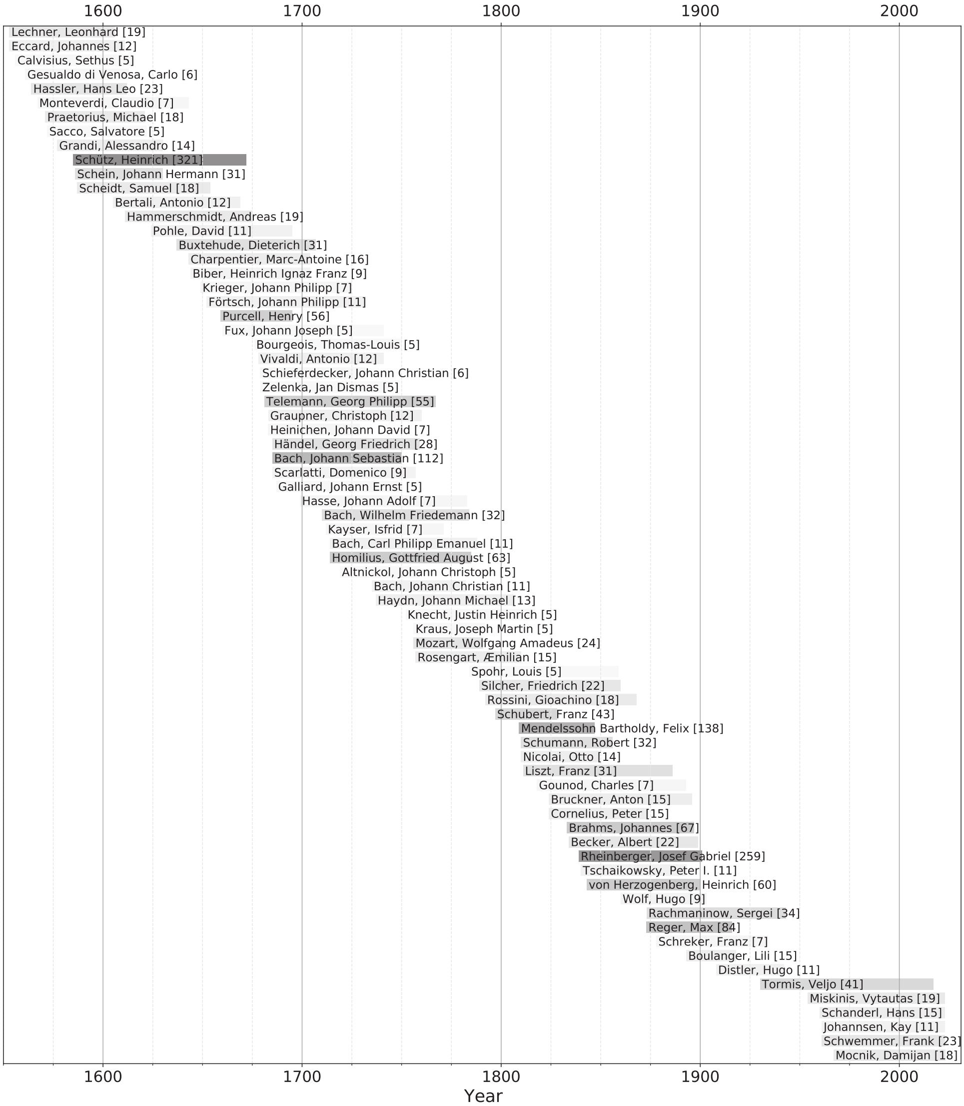


Fig. 3 Historical view of CAC considering all composers with at least five works. The number of works by each composer is indicated in square brackets and encoded by the darkness of the bars.

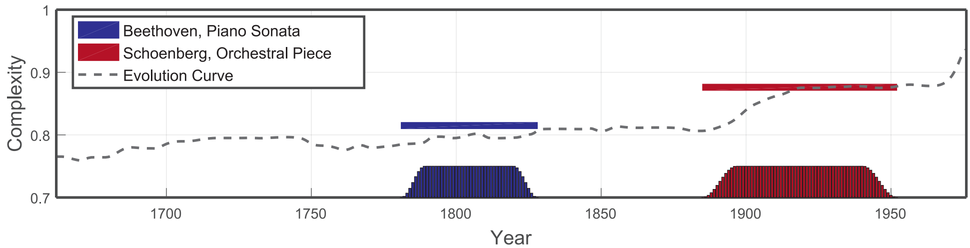


Fig. 4 Approximating evolution of tonal complexity based on composer dates, schematic example.

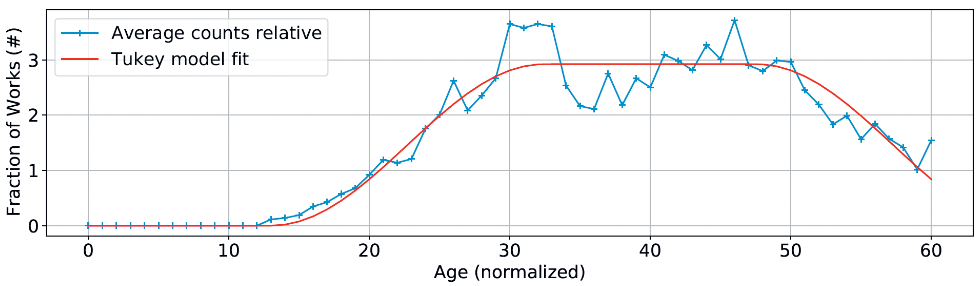


Fig. 5 Curve fitting procedure to determine the optimal window parameters.

puted a roughly flat distribution with smooth edges, using a so-called *Tukey window* (shown in Fig. 4). In Weiß et al. (2019), the parameters (start age and Tukey parameter α) were heuristically chosen. Since CAC contains such annotations for roughly half of the works (cf. Tab. 2), we can validate the approximation strategy and search for optimal values of the parameters. We do this in a stepwise fashion, by first determining the optimal start age, obtaining a value of 13. Then, we determine the optimal Tukey parameter to $\alpha = 0.72$. For details on the Tukey window and the fitting strategy, we refer to Weiß & Müller (2023).

The resulting curve is shown in Fig. 5 for an example hypothetic composer having died at the age of 60. With these optimized window parameters, we now validate the approximation strategy for the work count curve. To this end, we first compute the reference curve using the work dates for 1151 works that have these annotations. We post-process the curve with an average filter of length 15 years (red curve in Fig. 6). We then compare this reference curve with our approximation curve based on composer dates and our optimized Tukey window (blue curve in Fig. 6). Overall, the approximation seems to be suitable. Only in some periods (e.g., around 1680), the approximation curve is ahead, for others (e.g., at 1770), it lags behind the reference curve. We conclude that the approximation based on Tukey windows is a suitable strategy to compensate for missing work date annotations.

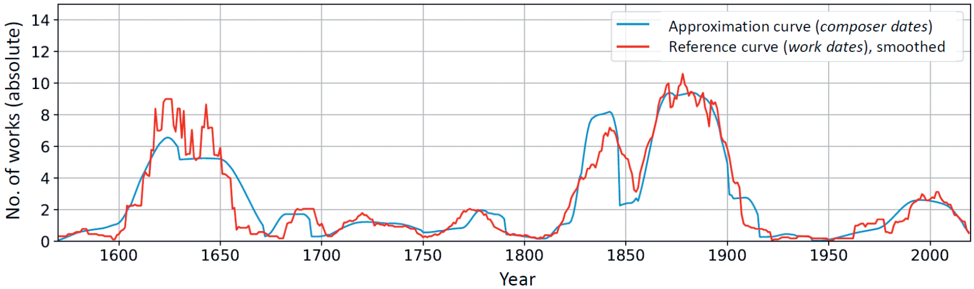


Fig. 6 Work count curves based on composer dates (approximation curve, blue) and based on work dates (reference curve, red), respectively.

3.3 Studying the Tonal Evolution of Sacred Music

With the presented strategy, we now investigate the tonal evolution of choral music in the CAC. To this end, we consider a computational approach for measuring tonal complexity from audio recordings (Weiß et al. 2019). Musical complexity is a highly relevant notion for analysis, which comprises several aspects such as acoustic, timbral, or rhythmic properties (Streich 2007). In our previous work (Weiß & Müller 2014), we introduced tonal complexity measures that locally describe distributions of energy across the twelve chromatic pitch classes used in the Western tonal system. Here, we considered the variety of pitch classes such that flat distributions (e.g., chromatic clusters) result in high complexity values while sharp distributions (e.g., single notes) result in low ones (see Fig. 7). To this end, we again rely on a chroma representation of the audio recording. For computing the complexity values, we map each chromagram frame (chroma vector) $c \in \mathbb{R}^{12}$ onto the circle of fifth. To this end, we first re-order the chroma values according to perfect fifth intervals. Based on the reordered vector, we compute circular statistics using the resultant vector. Then, our complexity measure $\Gamma(c) \in [0, 1]$ relates to the inverse length of this resultant vector. This measure describes the spread of the pitch classes around the circle of fifths, thus considering also the tonal relationship of active pitch classes. Fig. 7 illustrates the definition of the complexity measure and the resultant vector (in red) showing examples for three input chroma vectors. For a sparse vector (left), the complexity is minimal. For a flat vector (middle), we obtain maximal complexity. Other chroma vectors yield intermediate complexity values. We note that there are different strategies of aggregation to track-wise values. A local measure is obtained by calculating the complexity of each chromagram frame and then averaging over these values. A global chroma statistics can be computed by averaging chromagram frames first and then calculating a single complexity value for each track (movement). Aggregation to works is then done by averaging over the complexity values for all movements. In Fig. 6, we have studied the total number of works in CAC over the course of the years (work count curves)

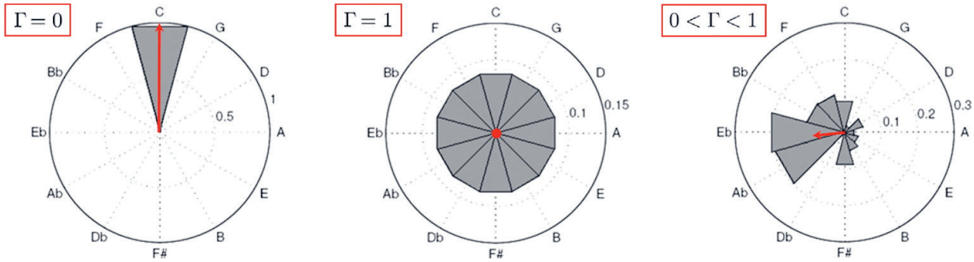


Fig. 7 Complexity measure Γ based on the circle of fifths. Values for a sparse chroma vector (left), a flat chroma vector (middle), and a more realistic chroma vector (right) are shown. The red arrows denote the resultant vectors.

using the work dates or our approximation strategy based on composer dates. We now apply these strategies to our measurements of tonal complexity. While the windows for each work were weighted with the value of 1 to account for the total number of works, we now use the complexity value of the respective work for weighting. We sum up all weighted windows and divide by the respective work count curve for normalization. We obtain a so-called *evolution* curve (EC) that indicates the average complexity of the works along the historical time axis. That way, each work contributes to the part of the time axis that corresponds to its work date (for the reference curve) or its composers' life dates (for the approximation curve).

We now apply this mixed strategy for investigating the evolution of the tonal complexity in CAC. Fig. 8 shows the resulting ECs, one based on the local and one on the global complexity. Looking at the global EC (black), we observe an increase in complexity over the course of the 17th and 18th century. Interestingly, we do not observe a drop around 1750, in contrast to (Weiß et al. 2019) where the demand for more "simplicity" after the Baroque era was clearly visible. On the other hand, the increase during the 19th century observed in Weiß et al. (2019) is not visible for CAC. Even more remarkable, CAC does not show any major increase in complexity during the 20th century. The modernism in tonality, pushed by expressionist and dodecaphonic composers such as A. Schoenberg or I. Stravinsky, does not seem to be reflected in choral music to the same degree. This could be based on different stylistics trends in choral music, but also be a property of the CAC, where complex atonal works might not be in the focus since they are hard to be performed by amateur choirs.

We finally show how the corpus analysis can be used for hypothesis-driven research, investigating the hypothesis that instrumental music is more complex than vocal music. We might expect such behavior since vocal compositions need to account for the higher difficulty in producing pitches when singing, especially for large and complex intervals. Moreover, musicologists usually claim that compositional "revolutions" were often happening in compact instrumental genres such as the string quartet. To test our hypothesis, we use the instrumentation annotations and

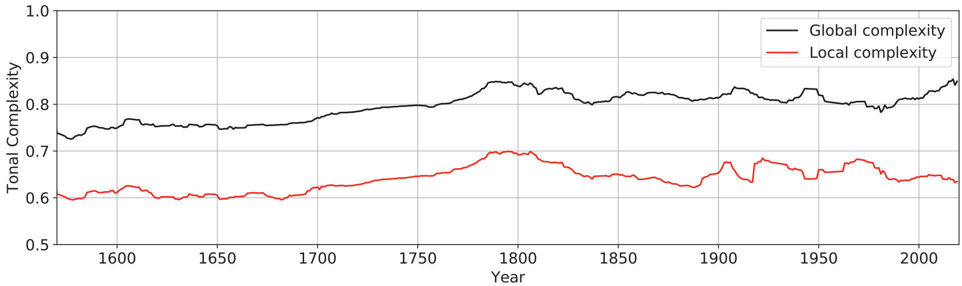


Fig. 8 Comparing ECs for global and local complexity.

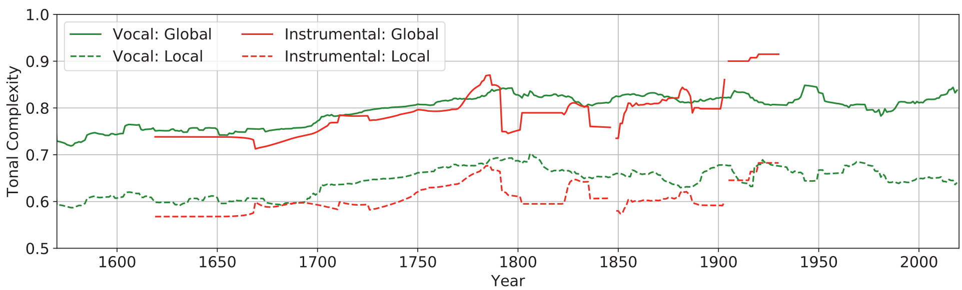


Fig. 9 Comparing ECs for global and local complexity separated into vocal and instrumental music.

compute a vocal as well as an instrumental evolution curve (Fig. 9). As a downside of CAC, we find an unbalanced situation (compare Tab. 2), resulting in a small number of works available for the instrumental EC. Nevertheless, we observe a clear tendency that contradicts our hypothesis: Vocal music seems to be more complex than instrumental music for most time periods. In particular, the offset is large for the local complexity (dashed lines). This may point to an interesting observation, but could also have a technical reason. Our chromagrams are based on a signal processing approach, which maps frequency components extracted from audio recordings to the twelve pitch classes. When dealing with recorded vocal music, this process often leads to substantial artifacts since pitch stability is much lower than for instruments and effects such as vibrato or portamento substantially blur the chromagrams. This may lead to quasi-chromatic artifacts that push the complexity measurements even locally. To overcome such issues, more recent chroma extraction strategies based on deep neural networks are very promising since they have shown to be successful for deriving tonal information from vocal recordings by reducing typical artifacts (Weiß & Peeters 2021).

4. Conclusions

These insights of corpus-based strategies demonstrate the high potential of audio recordings for research in computational theology. Summarizing this chapter, we emphasize the challenges that come along with the analysis of raw audio. Exact recognition of symbols (transcription) is hard and often infeasible. Nevertheless, there are computational approaches for deriving soft, probabilistic (“mid-level”) descriptions that strongly correlate with human understanding and intuition. In two case studies, we showed how such approaches may be employed to investigate large-scale musical works (as the *Johannespassion*) or whole corpora spanning several centuries (as for CAC). With the rapid development of processing techniques based on deep neural networks, a considerable improvement of such strategies can be expected in the near future. Nevertheless, in order to obtain reliable insights into theological and humanities question, an interdisciplinary dialogue between experts in both fields (theology and computer science) will remain essential.

Acknowledgments

To a large part, this chapter relies on previous work together with Meinard Müller and colleagues at the AudioLabs Erlangen. C.W. thanks the Carus-Verlag Stuttgart (Johannes Graulich and Ester Petri) for enabling the study of the Carus audio corpus.

References

Bäckström, T., Räsänen, O., Zewoudie, A., Zarazaga, P.P., Koivusalo, L., Das, S., Mellado, E.G., Mansali, M.B., Ramos, D., Kadiri, S., & Alku, P. (2022) Introduction to Speech Processing. 2. ed. [Computer Software]. *Zenodo.* DOI: https://doi.org/10.5281/zenodo.6821775 (Accessed: 21 June 2024).

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K., & Frommolt, K. (2010). Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring, *Pattern Recognition Letters,* 31, 1524–1534. DOI: https://doi.org/10.1016/j.patrec.2009.09.014 (Accessed: 21 June 2024).

Bellmann, H.G. (2012). *Categorization of Tonal Music Style. A Quantitative Investigation* [Diss.]. London: LAMBERT Academic Publishing.

Bello, J.P., Silva, C.T., Nov, O., DuBois, R.L., Arora, A., Salamon, J., Mydlarz, Ch., & Doraiswamy, H. (2019). SONYC. A system for monitoring, analyzing, and mitigating urban noise pollution, *Communications of the ACM,* 62, 68–77. DOI: https://doi.org/10.1145/3224204 (Accessed: 21 June 2024).

Benetos, E., Dixon, S., Duan, Z., & Ewert, S. (2019). Automatic music transcription. An overview, *IEEE Signal Processing Magazine,* 36, 20–30. DOI: https://doi.org/10.1109/MSP.2018.2869928 (Accessed: 21 June 2024).
Calvo-Zaragoza, J., Hajič Jr., J., & Pacha, A. (2020). Understanding optical music recognition, *ACM Computing Surveys,* 53(4), 1–35. DOI: https://doi.org/10.1145/3397499 (Accessed: 21 June 2024).
Mauch, M., MacCallum, R.M., Levy, M., & Leroi, A.M. (2015). The evolution of popular music. USA 1960–2010, *Royal Society Open Science,* 2, 1–10. DOI: https://doi.org/10.1098/rsos.150081 (Accessed: 21 June 2024).
Müller, M. (2021). *Fundamentals of Music Processing. Using Python and Jupyter Notebooks.* 2. ed. Cham: Springer. DOI: https://doi.org/10.1007/978-3-030-69808-9 (Accessed: 21 June 2024).
Nakamura, E., & Kaneko, K. (2019). Statistical evolutionary laws in music styles, *Scientific Reports,* 9, no pag. DOI: https://doi.org/10.1038/s41598-019-52380-6 (Accessed: 21 June 2024).
Scherbaum, F., Müller, M., & Rosenzweig, S. (2017). Analysis of the Tbilisi State Conservatory recordings of Artem Erkomaishvili in 1966. In *Proceedings of the 7th International Workshop on Folk Music Analysis* (FMA) (pp. 29–36). Málaga.
Smaragdis, P. (2004). Non-negative matrix factor deconvolution. Extraction of multiple sound sources from monophonic inputs. In C.G. Puntonet & A. Prieto (Eds.), *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation ICA* (pp. 494–499). Berlin/Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-540-30110-3_63 (Accessed: 21 June 2024).
Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec. Unsupervised pre-training for speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 3465–3469). Graz. DOI: https://doi.org/10.21437/Interspeech.2019-1873 (Accessed: 21 June 2024).
Streich, S. (2007) *Music Complexity. A Multi-Faceted Description of Audio Content* [Diss.]. Barcelona: University Pompeu Fabra.
Temperley, D. (1997). An algorithm for harmonic analysis, *Music Perception. An Interdisciplinary Journal,* 15, 31–68.
Weiß, Ch., Balke, S., Abeßer, J., & Müller, M. (2018). Computational corpus analysis. A case study on jazz solos. In *Proceedings of the 19th International Society for Music Information Retrieval Conference* (ISMIR) (pp. 416–423). Paris. DOI: https://doi.org/10.5281/zenodo.1492439 (Accessed: 21 June 2024).
Weiß, Ch., Mauch, M., Dixon, S., & Müller, M. (2019). Investigating style evolution of Western classical music. A computational approach, *Musicae Scientiae,* 23, 486–507. DOI: https://doi.org/10.1177/1029864918757595 (Accessed: 21 June 2024).
Weiß, Ch., & Müller, M. (2014). Quantifying and visualizing tonal complexity. In *Proceedings of the Conference on Interdisciplinary Musicology (CIM 14)* (pp. 184–187). Berlin. URL: https://www.audiolabs-erlangen.de/content/05_fau/

professor/00_mueller/03_ publications/2014_WeissMueller_TonalComplexity_CIM.pdf (Accessed: 21 June 2024).

Eid. (2021). Computergestützte Visualisierung von Tonalitätsverläufen in Musikaufnahmen. Möglichkeiten für die Korpusanalyse. In S. Klauk (Ed.), *Instrumentalmusik neben Haydn und Mozart. Analyse, Aufführungspraxis und Edition* (pp. 107–130). Würzburg: Königshausen & Neumann [= *Saarbrücker Studien zur Musikwissenschaft*, 20].

Eid. (2023). Studying tonal evolution of Western choral music. A corpus-based strategy. In A. Šela, F. Jannidis & I. Romanowska (Eds.), *Proceedings of the Computational Humanities Research Conference* (pp. 687–702). Paris. URL: https://ceur-ws.org/Vol-3558/paper7862.pdf (Accessed: 21 June 2024).

Weiß, Ch., & Peeters, G. (2021). Training deep pitch-class representations with a multi-label CTC loss. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)* (pp. 754–761). Online: Zenodo. DOI: https://doi.org/10.5281/zenodo.5624358 (Accessed: 21 June 2024).

White, Ch. W. (2013). *Some Statistical Properties of Tonality, 1650–1900* [Diss.]. New Haven, Connecticut: Yale University.

Figure Credits

Fig. 1–9 were generated by the author. Fig. 1 f. were published here for the first time. The others have already been published in previous publications as follows:

Fig. 3: Weiß & Müller 2023, 691.

Fig. 4: Weiß et al. 2019, 8.

Fig. 5: Weiß & Müller 2023, 694.

Fig. 6: Weiß & Müller 2023, 695.

Fig. 7: Weiß et al. 2018, 417.

Fig. 8: Weiß & Müller 2023, 699.

Fig. 9: Weiß & Müller 2023, 699.

Computational Tools and Methods for Film and Video Analysis

Manuel Burghardt[a], John Bateman[b], Eric Müller-Budack[c] and Ralph Ewerth[d]

[a] https://orcid.org/0000-0003-1354-9089, [b] https://orcid.org/0000-0002-7209-9295,
[c] https://orcid.org/0000-0002-6802-1241, [c] https://orcid.org/0000-0003-0918-6297

Abstract In this chapter, we provide an overview of computational tools and methods for film and video analysis. After introducing the evolution of methods in this field, we go on to lay some theoretical foundations for empirical video analyses. As we focus on using state of the art deep learning methods, we also provide an overview of the types of information that can already be extracted with these methods. Furthermore, we introduce an easy-to-use tool for video analytics, called TIB AV-Analytics (TIB AV-A). We demonstrate how TIB AV-A can be utilised to support the exploration of narrative patterns in the popular TV series *Game of Thrones.* Finally, we conclude by summarising the current state of available tools and methods for computational video analysis and outline some challenges that lie ahead.

Keywords computation film analysis, computational video analysis, distant viewing

1. A Short History of Computational Methods in Film and Videos Studies

Traditionally, the field of Digital Humanities (DH) has a strong focus on textual material, with its origins traced back to Roberto Busa's *Index Thomisticus.* However, in recent years, there has been a growing interest within DH towards film and video (Burghardt & Wolff 2016; Sittel 2017; Heftberger 2018; Burghardt et al. 2020; Arnold & Tilton 2022), leading to the establishment of dedicated special interest groups at both national[1] and international[2] levels. Given the highly interdisciplinary nature of DH, multiple views have emerged regarding the study of film. Burghardt et al. (2020) identify three key perspectives: (1) an *infrastructural perspective* encompassing GLAM (Galleries, Libraries, Archives, Museums) institutions and film archives, (2) a *media*

1 DHd AG Film & Video: https://dig-hum.de/ag-film-und-video (Accessed: 22 June 2024).
2 ADHO Special Interest Group AudioVisual material in Digital Humanities: https://avindhsig.wordpress.com/ (Accessed: 22 June 2024).

Manuel Burghardt, John Bateman, Eric Müller-Budack and Ralph Ewerth: Computational Tools and Methods for Film and Video Analysis, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 147–173. DOI: https://doi.org/10.11588/heibooks.1521.c21945

perspective addressing digital encounters in film and media studies, and (3) a *computational perspective,* which is focused on *multimedia information retrieval* and *multimodal information extraction.* As an increasing amount of video material is becoming available in digital form, the computational perspective has gained significant relevance, leading to the emergence of terms such as “distant viewing” (Arnold & Tilton 2019), “distant watching” (Howanitz 2015) and “deep watching” (Bermeitinger et al. 2019), to describe these developments. Wevers & Smits (2020) even propose a “visual digital turn”, driven by the capabilities of deep learning techniques.

While the conceptualisation of distant viewing and the like has gained recent attention, early examples of quantitative film studies with a focus on shot analyses can be found with Barry Salt (1974; 2006) and Yuri Tsivian’s *Cinemetrics* database (2009). Additionally, other quantitative approaches to stylistic and formal feature analysis have been explored. These include, for instance, language analysis (Hoyt et al. 2014; Byszuk 2020; Bednarek 2023) and colour analysis (Burghardt et al. 2018; Pause & Walkowski 2018; Flueckiger & Halter 2020). Another branch of quantitative analysis of large video corpora has its focus on information visualisation. Manovich’s (2013) *Visualizing Vertov* project represents an early example that heavily relies on visualisations to uncover patterns in extensive collections of images and videos. In addition to such visual analytics approaches, a wide range of tools is available for the annotation and analysis of videos and movies. Examples of these tools include *ELAN* (Wittenburg et al. 2006), *Videana* (Ewerth et al. 2009), *ANVIL* (Kipp 2014), and *VIAN* (Halter et al. 2019). For a comprehensive overview of these tools, along with their specific features and functionalities, we recommend the survey paper by Pustu-Iren et al. (2020). Furthermore, there are more recent tools that go beyond the analysis of visual aspects alone and encompass language (spoken and written) as well as audio (music and sound). Notable examples of such tools include Zoetrope (Tseng et al. 2023; Liebl & Burghardt 2023) and TIB AV-A (Springstein et al. 2023).

With this chapter we aim to provide an introduction to computer-assisted, empirical analysis of film and videos. In Section 2, we lay some theoretical foundations for such empirical analyses. Section 3 is about providing an overview of the types of information that can already be extracted using deep learning methods. As setting up such methods can be challenging without advanced technical knowledge, we also introduce an easy-to-use tool called TIB AV-Analytics (TIB AV-A, Section 4). Furthermore, we present a case study in Section 5, demonstrating how TIB AV-A can be utilised to support the exploration of narrative patterns in the popular TV series *Game of Thrones.* In Section 6, we conclude by summarising the current state of available tools and methods for computational video analysis and outline some challenges that lie ahead.

2. Theoretical foundations of empirical film analysis

The study of film has long employed a combination of qualitative and quantitative methodologies (Korte 2004). While qualitative approaches abound in the analysis of movies, as demonstrated by works such as Stam (2000), Prince (2007), Sikov (2010), Ryan & Lenos (2020), and many others, DH approaches to film studies tend to emphasise quantitative and empirical aspects. In this section, we aim to establish some theoretical foundations that contextualise empirical film studies further and relate such endeavours to broader analytic concerns.

A host of *external* engagements with film continue to be broadly relevant for, and may interact with, approaches to film within DH. These range from archiving, historical studies, production studies, investigations of the effects of developing display technologies, provision for information-retrieval from audiovisual data, reception studies (from psychological studies to reviews and critique), to cultural studies of cinema as an institution. In all cases, obtaining a tighter analytic hold *internally,* i.e., of film as artefacts with specific designs for aesthetic or other purposes, can be seen as an important step in understanding the medium. Such an analytic hold is arguably best achieved with the support of empirical studies whereby properties of films are successively revealed and used to draw ever deeper generalisations concerning their functioning. DH approaches to film thus typically seek research strategies that involve both quantitative studies of the distributions and patterning of measurable filmic properties and qualitative, more hermeneutic interpretations of those distributions and patterns (Flückiger 2011; Heftberger 2018).

The primary challenge to be faced in this context is then how to gain access to those details of highly complex audiovisual filmic artefacts that are relevant for interpretation and analysis. Since many questions raised concerning film are interpretative and hermeneutic in nature, it is by no means self-evident how quantitative approaches may support such concerns. This is, indeed, a very general philosophical point raised for many branches of the DH. In the case of film, early quantitative approaches drew much of their motivation from stylometry for literature, but were limited to manual methods restricting both the scale of studies and the kinds of features that could be considered. The most extensive programme of this kind, pursued for many years by Barry Salt (1974; 2007), consequently counted shot lengths and shot scales for selected portions of collections of films drawn from distinct periods and directors. Work in this tradition continues and has revealed a host of historical and regional developments; Cutting & Candan (2015), for example, report on an again predominantly manual analysis of shot lengths of 9,400 English-language films and 1,550 non-English language films released between 1912 and 2013. Similar broad historical changes have been reported for brightness, colour, and shot transitions (e.g., Cutting et al. 2011a; Cutting et al. 2011b). Redfern (2022b) sets out a host of practically applicable R-scripts (R Core Team 2016) that support these kinds of studies.

Heftberger (2018) pursues a different approach by relying on direct visualisations composed of still images taken from films, particularly focusing on the works of Dziga Vertov. Direct visualisations of perceptible features, such as brightness, colour, and so on, have been explored by several researchers in DH since the generation of such visual representations is relatively straightforward. It is, however, questionable to what extent relying simply on (typically) visual perception is an effective method for revealing deeper patterns of interest. In many respects, this is symptomatic of the current highly exploratory phase of DH film studies, where what is used for analysis is what is technologically feasible. To date, more extensive approaches that address a broader range of film phenomena remain limited. Bakels et al. (2020), for example, report on highly detailed, multitracked analyses of films that specifically target issues concerning the audiovisual construction of affect in films drawing on the AdA ontology of film analytic terms[3]. However, these are, again, still largely manual although automated digital techniques are now being successively added as well.

It is evident that empirical distributional studies benefit considerably from technological support so that time-consuming and error-prone manual analysis can be reduced to a minimum, but basic issues remain concerning the overall utility of such approaches to film. For further developments it will be important to relate the capabilities of computationally-supported analysis tools more closely to the research questions raised in the humanistic study of film. To date, there is still a gap to be bridged here. In Heftberger's case, motivation is provided by the fact that, as she notes, Vertov himself paid close attention to formal design features when constructing his films and so the formal analysis is well justified. It is less clear, however, that this can be adopted as a general guideline concerning how to go about film analysis empirically using computational and other quantitative methods.

One of the ways in which more general orientations are currently being developed draws much from a multimedia *information-retrieval* view of computational film analysis. Kurzhals et al. (2016), for example, seek to raise *visual movie analytics* to a more semantic level by combining a variety of sources of information (including scripts and subtitles) to answer the four basic questions of summarisation: *who, what, where,* and *when.* A rich range of automatic processing techniques are combined in this architecture to deliver descriptions of scenes and their events, thereby picking up from a quite different perspective some of the most traditional approaches to film analysis in terms of detailed shot-by-shot descriptions known as film or sequence protocols (cf. Kanzog 1991, 136–151; 163–183). Producing such protocols is extremely time consuming, however, and so any contribution to their automatic construction constitutes a useful advance; we will see many capabilities of these kinds emerging in the years to come and several significant developments in this direction are discussed below. Nevertheless, and more specifically for film as aesthetic artefacts, it is striking that the four basic summarisation questions omit one question that is crucial for film

3 See https://projectada.github.io/ontology (Accessed: 22 June 2024).

analysis: that is the *how* question. It is often insufficient to describe simply the bare plot structure of a narrative. For film we are often equally interested in the manner of the story telling since this is what drives film as a culturally effective communication form. The inclusion of such information was one of the primary reasons why traditional film protocols were so labour-intensive to produce.

This then returns us directly to several fundamental issues concerning the more interpretative nature of film analysis mentioned above and which DH still needs to come more to grips with. Since it is in general by no means straightforward to relate formal features of film with interpretations, principled ways of raising levels of abstraction to bridge the gap are needed. Vonderau (2017), for example, sets out a useful summary of some of the engagements between more digitally-oriented approaches to film as data on the one hand, and the traditional concerns of film studies on the other. But these questions raised anew within DH also have a longer history as a reoccurring point of criticism made by film scholars with respect to purely quantitative approaches in general. As David Bordwell and colleagues noted very early on, even though it is possible to compute "general statistical norms" in the style of Barry Salt, "such abstractions mean little without a concept of the range of paradigmatic choice dominant at a given period" (Bordwell et al. 1985, 60). Moreover, it is necessary to come to terms with an essential indirectness in the process of meaning attributions in the medium of film that is readily missed in quantitative approaches:

> Sometimes we're tempted to assign absolute meanings to angles, distances, and other qualities of framing. …The analysis of film as art would be a lot easier if technical qualities automatically possessed such hard-and-fast meanings, but individual films would thereby lose much of their uniqueness and richness. The fact is that framings have no absolute or general meanings. (Bordwell & Thompson 2008, 192)

It is only within particular systems of contrasts that specific meanings can be assigned at all (Branigan 1984, 29). Support for investigating this aspect of filmic meaning-making is only provided in current computational tools via interactive interfaces, since the computational processes themselves are not yet in a position to provide reliable hypotheses concerning interpretation. It is then, as we will see in subsequent sections, sensible to provide ready access to automatically captured formal features of film, but still leave the task of assigning interpretations to combinations of these features to human analysts by formulating appropriate queries by hand over combinations of automatically tagged categories (cf., e.g., Kurzhals et al. 2016).

Supporting interpretative tasks more directly will no doubt be a major area of development in the future. For this it will be useful to construct more robust theoretical frameworks that rely on explicit semiotic foundations more than has been the case in DH previously. Such foundations will need to align with contemporary views of semiotics that, on the one hand, are equally supportive of quantitative and qualitative kinds of description and, on the other, are capable of spanning both the recognition of more formal technical features and the entire process of hermeneutic interpretation in context. Only then can studies begin to balance the purely bottom-up, data-driven approaches that currently predominate (Redfern 2022b).

One such account of semiotics that has been related both to DH and to the analysis of film is set out in Bateman et al. (2017). In this approach expressive resources of a medium, which for film include montage, lighting, colour, music, and many more, are each characterised at three distinct levels of abstraction: *material* (supporting measurements), *form* (supporting explicit representations of paradigmatic systems of contrasts), and *discourse* (supporting interpretation in context). It is the latter level of description that moves the account beyond the capabilities of the more structural semiotics that became prevalent in the 1960s. The distinct levels of abstraction provided are directly relevant for DH because they motivate distinct classes of annotations that may be deployed to describe any artefacts being analysed and so serve to organise larger scale collections (Bateman 2022).

The multi-levelled view is now also essential for incorporating the new generation of computational techniques based on deep learning within a coherent overall framework. Such techniques are no longer limited to operating within single levels of abstraction and deliver useful results ranging from low-level formal features of film through to direct descriptions of semantic content. Semiotically, therefore, these components function similarly to the linguistic notion of constructions, which typically combine information from diverse levels of abstraction to offer reusable building-blocks for communication (Goldberg 1995). Specifically filmic constructions, often termed filmic idioms, are now also receiving formal treatments (Wu et al. 2018). The logical next step is therefore to combine these descriptions with automatic analysis components for prediction and recognition, and to support both visualisation and statistical evaluations building on the higher levels of abstraction achieved. Preliminary steps in this direction will be suggested below.

3. A short introduction to multimodal information extraction frameworks

Videos comprise multiple expressive modalities, such as image, audio (including speech), and text (overlaid text in video frames). To analyse individual videos and entire corpora, information from all modalities is required. Since the manual analysis of films is a very time-consuming task, there is a large need for automatic pattern recognition and multimedia retrieval methods to support DH researchers. In recent years, tremendous progress has been achieved in computer science fields such as computer vision, audio analysis, and natural language processing, thanks to deep learning models and the availability of large-scale datasets for training. In this section, we provide an overview of approaches for information extraction from images, audio, and text for film analysis, although it must be noted as well that this is only a selection of approaches made from those that we considered most relevant for video analysis, driven by our joint collaborations with researchers from the DH (i.e., film analysts, semioticians, media and communication scientists). In addition, there are further multimodal approaches that combine the capabilities of the methods listed here.

Computer vision: Different aspects of visual information are important for film analysis, ranging from low-level features (e.g., colour, brightness), over camera settings (e.g., shot scale, camera movement), to more complex information (e.g., actions, places, persons). There are many libraries (e.g., *scikit-learn*[4]) to extract low-level features such as *brightness, colour,* and *contrast.* For most other computer vision tasks, deep learning models such as convolutional neural networks (e.g., He et al. 2016) or transformers (e.g., Radford et al. 2021) are typically used. Methods for *temporal video segmentation* are an essential step to structure a video and can be categorised with regard to *shot* (e.g., Souček & Lokoč 2020), *scene, story, and topic boundary detection* (e.g., Wu et al. 2023). Another relevant information is the *camera setting* (e.g., shot scale, camera movement), camera pose (i.e., pitch, yaw, roll), and camera angle (e.g., Liu et al. 2022) that can be predicted by deep learning models. Approaches for *optical character recognition* (e.g., Kuang et al. 2021) automatically detect overlaid text in images that can be further analysed with approaches for natural language processing (see below). There are various deep learning approaches for image content analysis. In particular, the identification of *persons* (e.g., Deng et al. 2020), *facial attributes* (e.g., emotions, head pose, gender; Hempel et al. 2022; Serengil & Ozpinar 2021), and other *concepts* (e.g., animals, cars, objects; Radford et al. 2021) has been well-studied by the computer vision community. Furthermore, approaches for the identification of *place categories* (e.g., church, market, restaurant; Zhou et al., 2018), *geographical locations* (e.g., Müller-Budack et al. 2018; Theiner et al. 2022), and *events* (e.g., protests, elections, natural disasters; Müller-Budack et al. 2021) have been presented that can be used to categorise and characterise film segments. While deep learning models are often explicitly optimised for such tasks using labelled training data, recent vision-language models such as *CLIP* (*Contrastive Language-Image Pretraining;* Radford et al. 2021) have been trained with hundreds of millions of image-text pairs to implicitly learn visual concepts. These models can be applied to many tasks since they can measure the *similarity of arbitrary concepts* (e.g., objects, weather, occupation) to an image based on a textual description (i.e. a prompt). Recently, novel large vision-language models (e.g., Alayrac et al. 2022; Dai et al. 2023) combine the capabilities of these approaches with large language models such as OpenAI's GPT-4[5] (*Generative Pre-training Transformer* 4) and achieve impressive results for many applications,

4 See https://scikit-learn.org (Accessed: 22 June 2024).
5 See https://openai.com/gpt-4 (Accessed: 22 June 2024).

including film and video analysis (Zhang et al. 2023). While most aforementioned approaches focus on single images and have to be applied to each video frame, methods for video classification also consider temporal context from frame sequences (e.g., Ni et al. 2022) for further applications such as *action recognition* (e.g., running, talking).

Audio analysis: Basic analysis steps for audio regard *low-level features,* such as *amplitude, volume,* and *spectrogram* (e.g., using the *librosa* library for *Python*) that can indicate volume changes, (rhythmic) patterns, music, and other sound effects. The *transcription of the spoken language* is another highly relevant task for film analysis. Recently, neural transformer architectures have been introduced for automatic speech recognition (e.g., *Whisper;* Radford et al. 2023) that achieve impressive results across many languages.[6] Automatically extracted transcripts enable an in-depth analysis of speech using tools from natural language processing (see below). Methods for *speaker diarisation* (e.g. Bredin & Laurent, 2021) can further refine the speech transcript by assigning the identity of the corresponding speaker to, for example, analyse the spoken language for each speaker individually or to find forms of conversations (e.g., monologue, dialogue) in a film. It also serves as the basis for the *identification of voice characteristics,* such as gender (e.g., Baevski et al. 2020) and emotions (e.g., Ravanelli et al. 2021). Besides the analysis of speech, researchers have also focused on the *detection* and *classification of music* (e.g., Liu et al. 2021) as well as more general *audio classification* (e.g., Wu et al. 2022). Motivated by CLIP (see above), CLAP (*Contrastive Language-Audio Pretraining;* Wu et al. 2022) has been trained with several hundred-thousand audio and text pairs to enable classification of *arbitrary* audio concepts (e.g., sound events like *siren wailing* or *rain falling*) based on textual prompts.[7]

Natural language processing: As mentioned above, methods for optical character recognition from images (video frames) and automatic speech recognition from audio allow for the extraction of textual information from videos based on overlaid text and speech. Methods from natural language processing enable many perspectives to further work with such language data. For example, *part-of-speech tagging* (e.g., *spaCy*[8]) can be applied for syntax analysis, to better understand the grammatical structure of a sentence. *Named entity recognition* and *disambiguation* (e.g., *spaCy,* Wu et al. 2020) can automatically detect mentions of persons, locations, and events that play crucial roles in videos and films.[9] Moreover, there are numerous approaches for the classification of *topics* (e.g., Grootendorst 2022) and *sentiment* (e.g., Devlin et al. 2019) that can provide insights to the overall plot as well as the emotional tone and

6 See https://github.com/openai/whisper#available-models-and-languages (Accessed: 22 June 2024).
7 Cf. the chapter from Ch. Weiß in this volume.
8 See https://spacy.io (Accessed: 22 June 2024).
9 Cf. the chapter from E. Gius in this volume.

dynamics of characters throughout the film.[10] Very recently, large language models such as OpenAI's GPT-4 have been massively applied to a wide range of the aforementioned tasks and beyond.

4. Video analysis with the TIB AV-Analytics (TIB AV-A) Tool

The implementation of the deep learning techniques introduced in the previous section can pose substantial technical challenges. As an intermediate step, a number of toolkits[11] have been proposed that provide a basic layer of abstraction, but still require advanced technical knowledge and data literacy. However, to make available the advantages of large-scale pattern recognition and multimedia retrieval methods (see Section 3) to the broader community of scholars that work with audiovisual material, an easy-to-use tool with a graphical user interface is desirable. This is the main motivation for the TIB AV-Analytics platform (TIB AV-A[12]) that is currently being developed by TIB – Leibniz Information Centre for Science and Technology, in collaboration with film scholars from the University of Mainz.

TIB AV-A is a web-based platform for systematic film and video analysis (a screenshot is shown in Fig. 1). The platform uses modern web technologies and a plugin structure to simplify the integration of new plugins for developers and researchers to maintain TIB AV-A at the current state of the art. We use containers (e.g., *Docker*[13]) for virtualization for easy setup and to manage software dependencies, as well as an inference server (currently *Ray*[14]) for stable deployment. To ensure interoperability with other video analysis tools, TIB AV-A provides an Application Programming Interface (API) and import and export of results in common data formats, such as csv *(comma separated values)* files, as well as to the widely used ELAN video annotation tool (Wittenburg et al. 2006). The source code is publicly available.[15] More details are described by Springstein (2023).

In contrast to previous video analysis tools that either only allow for manual annotations (e.g., *ANVIL*[16], by Kipp 2014; *Cinemetrics*[17], by Tsivian 2009; *ELAN*[18], by

10 Cf. the chapters from M. Althage and R. Sprugnoli in this volume.

11 *Distant viewing toolkit,* Python notebooks (Arnold & Tilton 2020): https://github.com/distant-viewing/dvt; *Computational Film Analysis with R* (Redfern 2022b): https://cfa-with-r.netlify.app/index.html (Both accessed: 22 June 2024).

12 See https://service.tib.eu/tibava (Accessed: 22 June 2024).

13 See https://www.docker.com (Accessed: 22 June 2024).

14 See https://www.ray.io (Accessed: 22 June 2024).

15 See https://github.com/TIBHannover/tibava (Accessed: 22 June 2024).

16 See http://www.anvil-software.de (Accessed: 22 June 2024).

17 See https://cinemetrics.uchicago.edu (Accessed: 22 June 2024).

18 See https://archive.mpi.nl/tla/elan (Accessed: 22 June 2024).

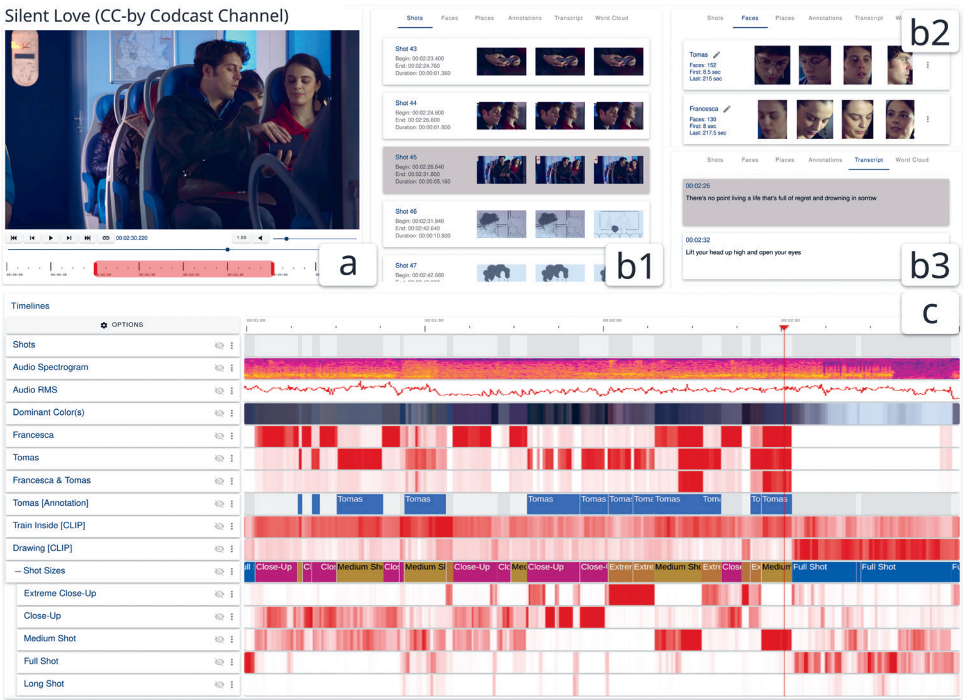


Fig. 1 Interface of TIB AV-A for the short movie *Silent Love* (CC-by Codcast Channel, Original Video: https://www.youtube.com/watch?v=KuuEs0oVVS8 [Accessed: 22 June 2024]). It contains a video player (a), an overview of detected shots (b1), persons (b2), and the speech transcript (b3). The timelines (c) can display categorical (e.g., *"Tomas"*) and numerical values (e.g., *"Drawing [CLIP]"*). Timelines with numerical values indicate, e.g., the probability whether a concept is depicted in a video. The user can select the visualisation type (line chart, colour chart) and colour (here: from white [unlikely] to red [likely]).

Wittenburg et al. 2006) or contain only a few selected methods for automatic content analysis (e.g., *Videana*, by Ewerth et al. 2009; *VIAN*[19], by Halter et al. 2019), TIB AV-A provides a vast collection of state-of-the-art pattern recognition approaches without the necessity of advanced technical knowledge or specific hardware requirements. Users from various disciplines can simply upload their own videos and then have access to a variety of analytical perspectives. An overview of currently supported methods for filmic analysis is provided in Tab. 1.

19 See https://www.vian.app (Accessed: 22 June 2024).

Tab. 1 Overview of current methods for image and video analysis as well as audio and speech analysis in TIB AV-A.

Image and video analysis	Basic image features: dominant **colour(s) and brightness**
	Shot boundary detection
	Cut frequency (cf. Redfern 2022a), i.e., the frequency of shot transitions
	Shot scale classification, to differentiate between the following shot scales: extreme close-ups, close-ups, medium shots, full shots, and long shots
	Place classification (e.g., church, market, restaurant, etc.)
	Person recognition based on an example image
	Place and person clustering to automatically find the most frequently appearing places and persons/actors
	Facial expression recognition (e.g., angry, happy)
	Zero-shot image classification for arbitrary visual concepts based on textual descriptions (e.g., "A photo taken in a train", see Fig. 1)
	Zero-shot video classification for arbitrary audiovisual concepts based on textual descriptions (e.g., "A video with celebrating people")
	Image captioning to automatically describe frames within a video
Audio and speech analysis	Basic audio features: **amplitude curve** (waveform), **volume** (root mean square), and the **frequency spectrum**
	Speech recognition to automatically transcribe speech in videos

Besides some standard analysis tasks (e.g., colour analysis, shot boundary detection), the addition of speech recognition and zero-shot image and video classification is most notable in TIB AV-A. High-quality transcripts (i.e., with a low word error rate) enable a much better analysis of speech using approaches from natural language processing for tasks like *topic modeling, named entity linking,* etc., which will be added in TIB AV-A in the future. Furthermore, zero-shot image and video classification enable various downstream tasks. Based on a textual prompt, the underlying vision language models, i.e., *CLIP* (Radford et al. 2021) and *InstructBLIP* (Dai et al. 2023), can recognise (a set of) *arbitrary* concepts. In this way, users can automatically search videos for various concepts ranging from real-world objects (e.g., flags, cars, etc.) and animals over environmental settings (e.g., places, weather, daytimes) to much more complex concepts, e.g., occupations of persons (e.g., police officer, reporter), events (e.g., natural disasters, demonstrations, types of sports), etc.

Although TIB AV-A provides a large set of state-of-the-art methods for automatic film analysis, DH researchers are often interested in more advanced patterns that can comprise a combination of features. For example, sequences in movies with high *shot density,* sudden *volume changes,* and *close-up shots* may indicate suspenseful key scenes or actions in movies. The combination of features can also add conditions to certain patterns to, for example, search for actions if a specific person or object is visible (see Fig. 1). To enable such combinations, TIB AV-A offers the option to aggregate probabilities of certain features (e.g., scenes, emotions, shot scales) with logical operations *(or/and)*. Based on the features extracted from a given video, users can create interactive visualisations for qualitative analysis. Currently, TIB AV-A supports a word cloud visualisation based on the extracted speech transcripts as well as scatter and line plots for which the user can display (and hide) specific features and feature combinations (see Fig. 1). Moreover, graphs that can, for example, show character constellations and their occurences at specific locations and places can be created.

5. Case study: Analysing the end of the *Game of Thrones* series for narrative patterns

We have seen in the previous section how state of the art tools for managing the automatic analysis of films provide support for a variety of analysis methods. The components being integrated into TIB AV-A cover two main kinds: first, the automatic analysis of films with respect to categories and properties that hold for all films, such as shot boundaries, colour ranges, sound spectrograms, and the like, and second, the automatic analysis of films with respect to categories, semantic constructs, or formal features that are selected by the human analyst. In both areas, we can expect the accuracy, precision, and diversity of results delivered to grow substantially in the coming years. Several questions remain, however, concerning how these capabilities can be leveraged to support the distinct kinds of analyses that may be targeted for film. In this section, we show an example of analysis that focuses specifically on uncovering larger-scale filmic structures that serve functional aims such as storytelling.

To make the discussion concrete, analysis will proceed with respect to the closing scenes of the last episode in the final series of *Game of Thrones*, created by David Benioff and D.B. Weiss for HBO and first aired in 2019.[20] The *what* of this segment of material is quickly described: the three main characters from the story's central Stark family, Jon Snow, Arya Stark, and Sansa Stark, begin new stages of their lives. Jon Snow passes the boundary separating civilisation from the icy north, Arya Stark sails west to look for new lands, and Sansa Stark is crowned queen. Thus the series ends. Filmically, however, the presentation of these events deploys a collection of well-known techniques yielding a tightly structured comparison of the respective fates of the individuals depicted. It becomes more relevant from the perspective of

20 The analysed scene can be seen at: https://www.youtube.com/watch?v=zUZvYAjaEZk (Accessed: 22 June 2024).

film analysis, therefore, to address the specifics of the *how* question concerning the segment's construction.

5.1 Workflow

We will show now how use of the TIB AV-A platform can support exploration of filmic organisation of this kind, revealing first the internal structure of the segment and then discussing briefly how this may be drawn into larger scale investigations of film form. We will also emphasise how working from the aesthetics and poetics of film analysis helps to set implementation priorities for the kinds of features that would be optimally beneficial for progressively moving ever more of the manual and semi-automatic analysis to automatic analysis. In the following workflow we also use the ELAN tool for manual annotation and correction of the automatic annotations and some custom R scripts for plotting of the results.

The first step is to load the film segment of interest into TIB AV-A and to perform the standard automatic processing pipelines of shot segmentation, shot scale, and so on. At this point, items that are known to be of particular relevance for the segment can also be used for specific categories – for example, searching for faces of the principal characters on the basis of uploaded images of their faces or by using natural language phrases for zero-shot content-based segmentation.

The second step is to export the analysis tracks from TIB AV-A and translate them into a form appropriate for further segmentation and manual annotation with ELAN or similar tools; this latter step is performed here locally using specific processing scripts. This allows errors in automatic processing to be corrected and further filmic features to be added that are not yet provided automatically by TIB AV-A. Relevant examples of these in the present case are camera movements since the segment relies extensively on camera movement cohesion across subsequences. The general scheme of analysis then follows that set out in Bateman & Schmidt (2012), where shots are allocated to spatiotemporal regions. Human visual perception is generally very fast and accurate in deciding whether it has encountered a particular place before and this kind of continuity is well-known from psychological studies as a fundamental unit for extended discourse comprehension (Zacks 2010; Loschky et al. 2020). Annotation tracks, or tiers, are consequently defined in ELAN so that shots can be assigned to them as revealed by any of the levels of analysis available. This is an area where increasingly accurate scene recognition combined with visual similarity measures can be expected to provide substantial improvements in the near future for supporting automatic or semi-automatic analysis. By these means we can see how questions driven directly by the needs of film poetic and aesthetic analysis may be progressively taken over and supported by the developing computational tools; required features that are missing may first be added manually and then supported computationally as they become available.

The third, and for current purposes final, step is to export the ELAN analysis further for focused examination of reoccurring filmic patterns. For this we use custom-built R scripts running locally that directly transform ELAN annotations into visualisations of the filmic structure, overlaid with results of automatic and manual analysis as desired. Whereas many classical formal editing features can now be explored in rather sophisticated ways for their statistical properties in R (cf. Redfern 2022b), here we will be concerned more with deriving higher-level organisational properties that often correspond more directly with interpretations. The visualisations employed here are defined in Bateman & Schmidt (2012) and draw loosely on musical notation, laying out successive shots horizontally so that further structural relationships, properties and groupings can be added freely. In short, we attempt here to identify functionally relevant sequences of combinations of filmic features that can start moving us beyond overall statistics of transitions, co-occurrences, and the like (Bateman 2014).

5.2 Analysis

The basic structure of the example segment is then as given in Fig. 2. This shows the shots of the scene running horizontally numbered along the bottom row, together with brief functional descriptions of those shots included for ease of reference along the top row. Whenever what has been identified as a shot responds to distinct functional groupings, it is further divided into 'subshots' – as seen, for example, in shot 15, which further divides into a segment tracking a character walking (15.1) followed by a stationary focus on that character (15.2).

This visualisation readily allows us to see the essentially three-line development of the sequence, where successive shots frequently range across the distinct locales of the three main characters (arrayed vertically). This structure is defined formally in Bateman & Schmidt (2012, 222–226) as a tri-partitioned polyspatial alternation and commonly expresses contrast and comparison. Each shot is also labelled here with its shot scale, running from tight close-up or detail shots (TS) to extra long shots (ELS). The sequence thus starts with three tight shots running successively across Jon Snow's (JS), Arya Stark's (AS), and Sansa Stark's (SS) locales; the next three shots, also tight shots, repeat these transitions in reversed order; and so on.

Filmically it is then interesting to examine more closely how the construction of the segment maintains coherence despite this rapid cross-cutting between scenes. To explore this, we successively augment the visualisation with further layers of information from the annotation. Fig. 3, for example, shows the visualisation with the annotation tracks of distinct kinds of camera use folded in, both as labels and as coloured groupings shown over the affected shots. The shots classified according to *camera direction* in this figure then show well how direction maintains cohesion across the multiple locales. Shots 18–22, for example, maintain rightwards tracking,

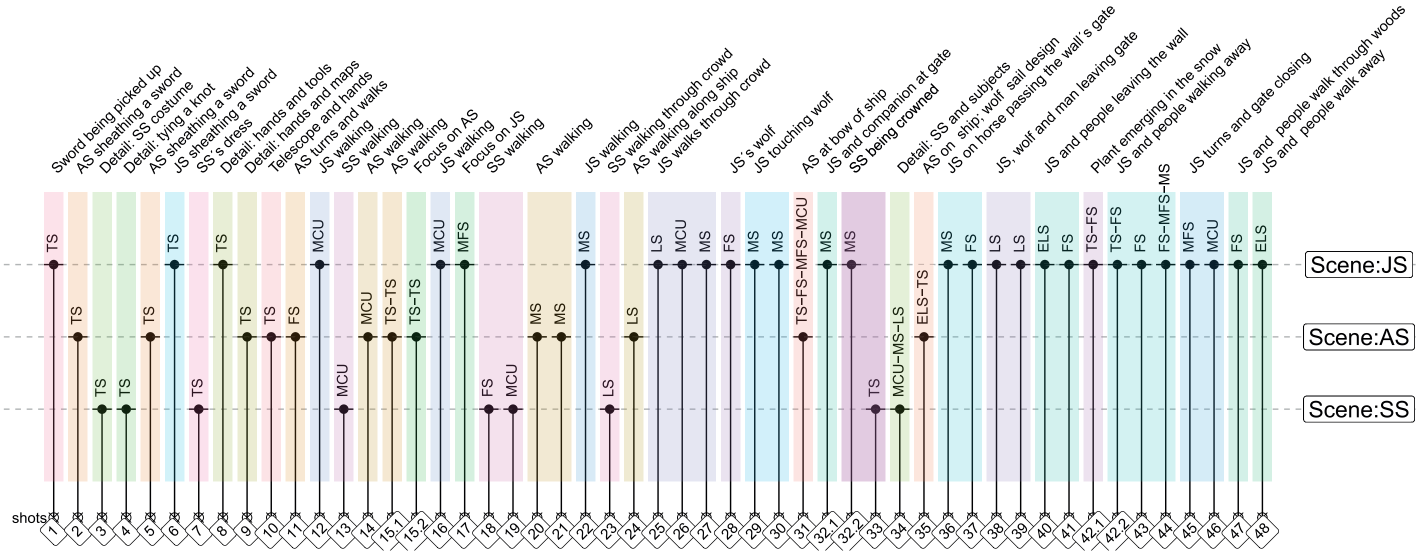


Fig. 2 Basic visualisation of the annotated film structure of the *Game of Thrones* segment; shots run horizontally, numbered at the bottom; colouring indicates semantic content grouping (all graphs produced with the R-package *ggplot*, Wickham 2016). The shot-scale abbreviations are based on standard shot scales, in increasing distance: *tight or detail shot* (TS), *closeup* (CU), *medium closeup* (MCU), *full shot* (FS), *medium full shot* (MFS), *medium shot* (MS), *long shot* (LS), *extra long shot* (ELS).

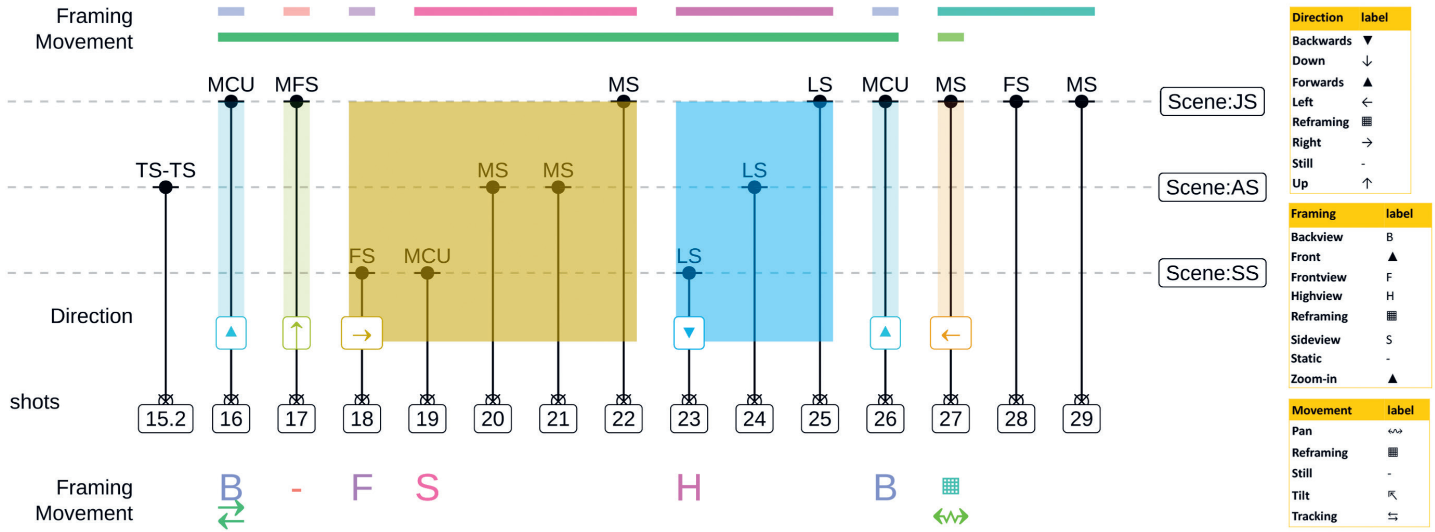


Fig. 3 Visualisation of the annotated film structure of the *Game of Thrones* segment (shots 15 – 29) augmented with camera use information. In this visualisation, camera direction has been prioritised showing grouping with larger coloured blocks. The bars along the top show the grouping that is imposed by framing and movement; the symbols at the bottom show what kind of camera use is involved in each case.

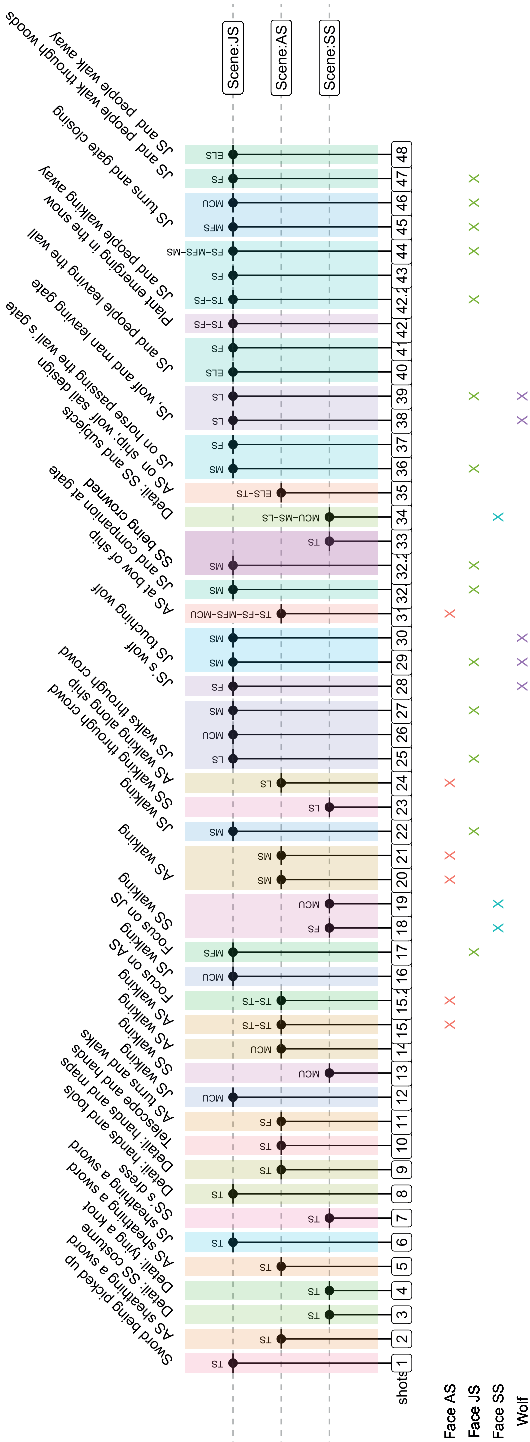


Fig. 4 Visualisation of the annotated film structure of the *Game of Thrones* segment with the results of automatic face recognition and CLIP-based semantic concept recognition.

while shots 23–25 maintain the camera moving out of the scene space. Successive shots within the same locale (shots 28–30, 38–40, 42–46 in Fig. 2), in contrast, appear without any marked camera direction. And, crucially, none of these filmic technical features carries the meaning of the construction alone; it is only in their structural composition that a reliably interpretable form emerges.

The coloured bars running across the top of the graph show similarly how other dimensions of camera use, here framing and movement, also serve to group shots, again often across the three locales. The lower symbol lines of the graph indicate which kinds of framing and movement are playing out more finely; shots 19–22, for example, all exhibit a constant framing (S: "sideview") that contributes to holding the sequence together. Any of these filmic properties can be selected for visual prominence in the graphs so that the differing kinds of structures can be made visible. Relevant examples here would be continuous diegetic sound matching across scenes (such as footsteps) as well as the overall non-diegetic music organising the segment, with the *Stark family motif* running through shots 1–17, gradually blending with the *Game of Thrones* theme across shots 18–30, which then takes over in the remaining shots 31–48.

It is also possible to overlay other automatic annotation results obtained from TIB AV-A. Fig. 4, for example, shows where TIB AV-A recorded one of the three main faces of the characters occurring with high confidence (classified by visual similarity) and occurrences of Jon Snow's wolf (classified by zero-shot semantic classification using CLIP as described above). Here it is interesting for the filmic construction of the sequence that identification of the protagonists is left quite late: only from shot 15 onwards are faces shown as the story moves towards its finale. The placement of the wolf in the middle of the scenes involving Jon Snow is also quite accurate.

While the views shown so far support further exploration of how this sequence is constructed, for the future it will be beneficial to define structural patterns on the basis of the revealed structures that can then in turn be fed back into the automatic search capabilities of TIB AV-A and other tools. This requires the definition of patterns as search queries. For the present case, for example, one would want to search for repeating sequences of shots that are each drawn from a different location but which nevertheless maintain a collection of identical formal technical features, such as camera movement, direction, and so on. Extending such pattern queries to include any of the possible automatically ascertained features promises to dramatically change the state of the art for computer-supported film analysis at scale as well as renewing contact with more hermeneutically-driven research challenges.

6. Conclusion and challenges ahead

We have come a long way in the analysis of film and video in DH. In the era of deep learning advancements, an extensive array of methods has emerged, facilitating the automated extraction of diverse multimodal features. This surge in quantitative data availability necessitates the development of a corresponding analytical framework. We believe that such a framework should be grounded both in empirical standards and in theoretical underpinnings such as multimodal theory and semiotics. It would also benefit from integrating concepts from common taxonomies used by researchers in the DH, such as the AdA film ontology[21], for which first promising experiments have been conducted with TIB AV-A. However, a suitable integration that also captures the hierarchical nature of such ontologies remains to be developed.

While empirical analysis can be seen as a cornerstone of computational film analysis, there's a compelling argument for the integration of exploratory tools like TIB AV-A. So far, most existing tools focus on the exploration and visualisation of single videos. To actually realise the concept of distant viewing (Arnold & Tilton 2019) across multiple videos, we must devise methods for simultaneously visualising multiple videos – an intricate task due to the dynamic nature inherent to video content. First strides have already been taken in this direction through cultural analytics (Manovich 2020) and visual movie analytics (Kurzhals et al. 2016). However, as the trend in textual DH goes towards scalable viewing (Weitin 2017), i.e. a hybrid approach that allows scholars to transition smoothly between close and distant viewing perspectives, this concept holds promise for the analysis of video material as well. Some first examples of scalable viewing can be found in approaches for the visualisation of news videos (Liebl & Burghardt 2023b; Ruth et al. 2023) as well as more generic tools such as *PixPlot*[22] or the *Collection Space Navigator* (Ohm et al. 2023).

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barrerira, R., Vinyalis, O., Zisserman, A., & Simonyan, K. (2022). Flamingo. A Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems 35* (pp. 23716–23736). New Orleans, Louisiana: Neural Information Processing Systems. DOI: https://doi.org/10.48550/arXiv.2204.14198 (Accessed: 22 June 2024).

21 See https://projectada.github.io/ontology (Accessed: 22 June 2024).

22 *PixPlot*, from Yale DH Lab: https://github.com/YaleDHLab/pix-plot (Accessed: 22 June 2024).

Arnold, T., & Tilton, L. (2019). Distant viewing. Analyzing large visual corpora, *Digital Scholarship in the Humanities,* 34(1), 3–16. DOI: https://doi.org/10.1093/llc/fqz013 (Accessed: 22 June 2024).

Eid. (2020). Distant Viewing Toolkit. A Python Package for the Analysis of Visual Culture, *Journal of Open Source Software,* 5(45).

Eid. (2022). Analyzing Audio/Visual Data in the Digital Humanities. In J. O'Sullivan (Ed.), *The Bloomsbury Handbook to the Digital Humanities* (pp. 179–187). London: Bloomsbury Publishing.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0. A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems 33* (pp. 12449–12460). Online: Neural Information Processing Systems. DOI: https://doi.org/10.48550/arXiv.2006.11477 (Accessed: 22 June 2024).

Bakels, J.-H., Grotkopp, M., Scherer, T., & Stratil, J. (2020). Digitale Empirie? Computergestützte Filmanalyse im Spannungsfeld von Datenmodellen und Gestalttheorie, *Montage AV – Zeitschrift Für Theorie Und Geschichte Audiovisueller Kommunikation,* 29(1), 99–118.

Bateman, J. A. (2014). Looking for what counts in film analysis. A programme of empirical research. In D. Machin (Ed.), *Visual Communication* (pp. 301–330). Berlin/Boston: De Gruyter Mouton.

Id. (2022). Growing theory for practice. Empirical multimodality beyond the case study, *Multimodal Communication,* 11(1), 63–74. DOI: https://doi.org/10.1515/mc-2021–0006 (Accessed: 22 June 2024).

Id., & Schmidt, K.-H. (2012). *Multimodal Film Analysis. How Films Mean.* London: Routledge.

Bateman, J., Wildfeuer, J., & Hiippala, T. (2017). *Multimodality. Foundations, Research and Analysis. A Problem-Oriented Introduction.* Berlin/Boston: De Gruyter Mouton.

Bednarek, M. (2023). *Language and Characterisation in Television Series. A corpus-informed approach to the construction of social identity in the media.* Amsterdam: John Benjamins Publishing Company [= *Studies in Corpus Linguistics,* 106].

Bermeitinger, B., Gassner, S., Handschuh, S., Howanitz, G., Radisch, E., & Rehbein, M. (2019). Deep Watching. Towards New Methods of Analyzing Visual Media in Cultural Studies. In *Book of Abstracts of the International Digital Humanities Conference (DH).* Utrecht: Alliance of Digital Humanities Organizations. DOI: https://doi.org/10.13140/RG.2.2.12763.72486 (Accessed: 22 June 2024).

Bordwell, D., & Thompson, K. (2008). *Film Art. An Introduction.* New York: McGraw Hill.

Eid., & Staiger, J. (1985). *The Classical Hollywood Cinema. Film, Style and Mode of Production to 1960.* New York: Columbia University Press.

Branigan, E. (1984). *Point of View in the Cinema.* Berlin/Boston: De Gruyter Mouton.

Bredin, H., & Laurent, A. (2021). End-To-End Speaker Segmentation for Overlap-Aware Resegmentation. In *Proceedings of the Interspeech 2021* (pp. 3111–3115). Brno: International Speech Communication Association. DOI: https://doi.org/10.21437/Interspeech.2021-560 (Accessed: 22 June 2024).

Burghardt, M., Heftberger, A., Pause, J., Walkowski, N.-O., & Zeppelzauer, M. (2020). Film and Video Analysis in the Digital Humanities. An Interdisciplinary Dialog, *Digital Humanities Quarterly,* 14(4), 1–37. URL: http://www.digitalhumanities.org/dhq/vol/14/4/000532/000532.html (Accessed: 22 June 2024).

Burghardt, M., Kao, M., & Walkowski, N.-O. (2018). Scalable MovieBarcodes. An Exploratory Interface for the Analysis of Movies. In *Vis4DH. 3rd IEEE VIS Workshop on Visualization for the Digital Humanities.* Berlin: Institute of Electrical and Electronics.

Burghardt, M., & Wolff, Ch. (2016). Digital Humanities in Bewegung. Ansätze für die computergestützte Filmanalyse. In E. Burr (Ed.), *DHd 2016. Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts.* 2. überarbeitete und erweiterte Auflage (pp. 191–195). Leipzig: Verband Digital Humanities im deutschsprachigen Raum. URL: https://www.dhd2016.de/sites/default/files/dhd2016/files/boa-2.0_ohne_Vorwort.pdf (Accessed: 22 June 2024).

Byszuk, J. (2020). The Voices of Doctor Who. How Stylometry Can be Useful in Revealing New Information About TV Series, *Digital Humanities Quarterly,* 14(4). URL: http://www.digitalhumanities.org/dhq/vol/14/4/000499/000499.html (Accessed: 22 June 2024).

Cutting, J. E., Brunick, K. L., & DeLong, J. E. (2011a). The changing poetics of the dissolve in Hollywood film, *Empirical Studies of the Arts,* 29(2), 149–169.

Eid., Iricinschi, C., & Candan, A. (2011b). Quicker, faster, darker. Changes in Hollywood film over 75 years, *I-Perception,* 2(6), 569–576.

Cutting, J. E., & Candan, A. (2015). Shot Durations, Shot Classes, and the Increased Pace of Popular Movies, *Projections. The Journal for Movies and Mind,* 9(2), 40–62.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. (2023). InstructBLIP. Towards General-purpose Vision-Language Models with Instruction Tuning [Preprint]. *arXiv.* DOI: https://doi.org/10.48550/arXiv.2305.06500 (Accessed: 22 June 2024).

Deng, J., Guo, J., Liu, T., Gong, M., & Zafeiriou, S. (2020). Sub-center arcface. Boosting face recognition by large-scale noisy web faces. In A. Vedaldi, H. Bischof, T. Brox, & J. M. Frahm (Eds.). *Proceedings of the European Conference on Computer Vision 2020* (pp. 741–757). Cham: Springer [= *Lecture Notes in Computer Science,* 12356]. DOI: https://doi.org/10.1007/978-3-030-58621-8_43 (Accessed: 22 June 2024).

Devlin, J., Chang, M.-W., Kenton, L., & Toutanova, K. (2019). BERT. Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the*

Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies 2019 (pp. 4171–4186). Minneapolis: Association for Computational Linguistics. DOI: https://doi.org/10.18653/V1/N19-1423 (Accessed: 22 June 2024).

Ewerth, R., Mühling, M., Stadelmann, T., Gllavata, J., Grauer, M., & Freisleben, B. (2009). Videana. A Software Toolkit for Scientific Film Studies. In M. Ross, M. Grauer & B. Freisleben (Eds.), *Digital Tools in Media Studies. Analysis and Research. An Overview* (pp. 101–116). Bielefeld: Transcript Verlag.

Flückiger, B. (2011). Die Vermessung ästhetischer Erscheinungen, *Zeitschrift für Medienwissenschaft,* 3(2), 44–60. DOI: https://doi.org/10.1524/zfmw.2011.0022 (Accessed: 22 June 2024).

Ead., & Halter, G. (2020). Methods and Advanced Tools for the Analysis of Film Colors in Digital Humanities, *Digital Humanities Quarterly,* 14(4), 1–115. URL: http://www.digitalhumanities.org/dhq/vol/14/4/000500/000500.html (Accessed: 22 June 2024).

Goldberg, A. E. (1995). *Constructions. A construction grammar approach to argument structure.* Chicago: University of Chicago Press.

Grootendorst, M. (2022). BERTopic. Neural topic modeling with a class-based TF-IDF procedure. *arXiv.* DOI: https://doi.org/10.48550/arXiv.2203.05794 (Accessed: 22 June 2024).

Halter, G., Ballester-Ripoll, R., Flueckiger, B., & Pajarola, R. (2019). VIAN. A Visual Annotation Tool for Film Analysis, *Computer Graphics Forum,* 38(3), 119–129. DOI: https://doi.org/https://doi.org/10.1111/cgf.13676 (Accessed: 22 June 2024).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016* (pp. 770–778). Las Vegas: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.1109/CVPR.2016.90 (Accessed: 22 June 2024).

Heftberger, A. (2018). *Digital Humanities and Film Studies. Visualising Dziga Vertov's Work.* Basel: Springer International Publishing.

Hempel, T., Abdelrahman, A. A., & Al-Hamadi, A. (2022). 6d Rotation Representation For Unconstrained Head Pose Estimation. In *Proceedings of the IEEE International Conference on Image Processing 2022* (pp. 2496–2500). Bordeaux: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.48550/arXiv.2202.12555 (Accessed: 22 June 2024).

Howanitz, G. (2015). Distant Waching. Ein quantitativer Zugang zu YouTube-Videos. In *DHd 2015. Von Daten zu Erkenntnissen. Book of Abstracts* (pp. 33–38). Graz: Verband Digital Humanities im deutschsprachigen Raum. URL: https://gams.uni-graz.at/o:dhd2015.abstracts-gesamt (Accessed: 22 June 2024).

Hoyt, E., Ponto, K., & Roy, C. (2014). Visualizing and Analyzing the Hollywood Screenplay with ScripThreads, *Digital Humanities Quarterly,* 8(4), 1–57. URL: http://www.digitalhumanities.org/dhqdev/vol/8/4/000190/000190.html (Accessed: 22 June 2024).

Huang, Q., Xiong, Y., Rao, A., Wang, J., & Lin, D. (2020). MovieNet. A Holistic Dataset for Movie Understanding. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Proceedings of the European Conference on Computer Vision 2020* (pp. 709–727). arXiv: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.48550/arXiv.2007.10937 (Accessed: 22 June 2024).

Kanzog, K. (1991). *Einführung in die Filmphilologie.* München: Diskurs Film.

Kipp, M. (2014). ANVIL: The Video Annotation Research Tool. In J. Durand, U. Gut & G. Kristoffersen (Eds.), *The Oxford Handbook of Corpus Phonology* (pp. 420–436). Oxford: Oxford University Press.

Korte, H. (2004). *Einführung in die systematische Filmanalyse. Ein Arbeitsbuch.* 3. ed. Berlin: Erich Schmidt Verlag.

Kuang, Z., Sun, H., Li, Z., Yue, X., Lin, T.H., Chen, J., Wei, H., Zhu, Y., Gao, T., Zhang, W., Chen, K., Zhang, W., & Lin, D. (2021). MMOCR. A Comprehensive Toolbox for Text Detection, Recognition and Understanding. In *Proceedings of the 29th ACM International Conference on Multimedia 2021* (pp. 3791–3794). arXiv: Association for Computing Machinery. DOI: https://doi.org/10.48550/arXiv.2108.06543 (Accessed: 22 June 2024).

Kurzhals, K., John, M., Heimerl, F., Kuznecov, P., & Weiskopf, D. (2016). Visual Movie Analytics, *IEEE Transactions on Multimedia,* 18(11), 2149–2160. DOI: https://doi.org/10.1109/TMM.2016.2614184 (Accessed: 22 June 2024).

Liebl, Ch., & Burghardt, M. (2023). Zoetrope. Interactive Feature Exploration in News Videos. In W. Scholger, G. Vogeler, T. Tasovac, A. Baillot, & P. Helling (Eds.), *Digital Humanities 2023. Collaboration as Opportunity* (pp. 432–434). Graz: Alliance of Digital Humanities Organisations. DOI: https://doi.org/10.5281/zenodo.7961822 (Accessed: 22 June 2024).

Liu, C., Feng, L., Liu, G., Wang, H., & Liu, S. (2021). Bottom-up broadcast neural network for music genre classification, *Multimedia Tools and Applications,* 80(5), 7313–7331. DOI: https://doi.org/10.48550/arXiv.1901.08928 (Accessed: 22 June 2024).

Liu, S., Nie, X., & Hamid, R. (2022). Depth-Guided Sparse Structure-from-Motion for Movies and TV Shows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022* (pp. 15980–15989). New Orleans, LA: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.48550/arXiv.2204.02509 (Accessed: 22 June 2024).

Loschky, L.C., Larson, A.M., Magliano, J.P., & Smith, T.J. (2015). What would Jaws do? The tyranny of film and the relationship between gaze and higher-level narrative film comprehension, *PloS one,* 10(11), 1–23. DOI: https://doi.org/10.1371/journal.pone.0142474 (Accessed: 22 June 2024).

Manovich, L. (2013). Visualizing Vertov, *Russian Journal of Communication,* 5(1), 44–55.

Id. (2020). *Cultural Analytics.* Cambridge, Mass.: MIT Press.

Monaco, J. (2009). *How to Read a Film. Movies, Media and Beyond.* Oxford: Oxford University Press.

Müller-Budack, E., Pustu-Iren, K., & Ewerth, R. (2018). Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification. In Ferrari, V., Hebert, M., Sminchisescu, C., & Weiss, Y. (Eds.), *Computer Vision. ECCV 2018* (pp. 575–592). Springer, Cham [= *Lecture Notes in Computer Science,* 11216]. DOI: https://doi.org/10.1007/978-3-030-01258-8_35 (Accessed: 22 June 2024).

Müller-Budack, E., Springstein, M., Hakimov, S., Mrutzek, K., & Ewerth, R. (2021). Ontology-driven event type classification in images. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision 2021* (pp. 2928–2938), Waikoloa: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.48550/arXiv.2011.04714 (Accessed: 22 June 2024).

Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., & Ling, H. (2022). Expanding Language-Image Pretrained Models for General Video Recognition. In Avidan, Sh., Brostow, G., Moustapha, C., Farinella, G.M., & Hassner, T. (Eds.), *Proceedings of the European Conference on Computer Vision 2022* (pp. 1–18). Cham: Springer [= *Lecture Notes in Computer Science,* 13664].

Ohm, T., Solà, M.C., Karjus, A. & Schich, M. (2023). Collection Space Navigator. An Interactive Visualization Interface for Multidimensional Datasets, *arXiv.* DOI: https://doi.org/10.48550/arXiv.2305.06809 (Accessed: 22 June 2024).

Prinz, S. (2007). *Movies and meaning. An introduction to film.* 4. ed. Boston: Allyn & Bacon.

Pustu-Iren, K., Sittel, J., Mauer, R., Bulgakowa, O., & Ewerth, R. (2020). Automated Visual Content Analysis for Film Studies. Current Status and Challenges, *Digital Humanities Quarterly,* 14(4), 1–102. URL: http://www.digitalhumanities.org/dhq/vol/14/4/000518/000518.html (Accessed: 22 June 2024).

R-Kernteam. (2016). R. Eine Sprache und Umgebung für statistische Berechnungen [Computersoftware]. *R Foundation for Statistical Computing.* URL: https://www.R-project.org (Accessed: 22 June 2024).

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In M. Meila & T. Zhang (Eds.), *Proceedings of the International Conference on Machine Learning* (pp. 8748–8763). arXiv. [= *Proceedings of Machine Learning Research,* 139]. DOI: https://doi.org/10.48550/arXiv.2103.00020 (Accessed: 22 June 2024).

Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the International Conference on Machine Learning 2023* (pp. 28492–28518). Honolulu: International Machine Learning Society [= *Proceedings of Machine Learning Research,* 202]. DOI: https://doi.org/10.48550/arXiv.2212.04356 (Accessed: 22 June 2024).

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., De Mori, R., Bengio, Y. (2021). SpeechBrain. A General-Purpose Speech Toolkit [Preprint]. *arXiv.* DOI: https://doi.org/10.48550/arXiv.2106.04624 (Accessed: 22 June 2024).

Redfern, N. (2022a). Analysing Motion Picture Cutting Rates, *Wide Screen,* 9(1). 1–29. URL: https://widescreenjournal.org/vol-9-no-1-2022-title (Accessed: 22 June 2024).

Id. (2022b). *Computational Film Analysis with R.* Version 0.9.004. Zenodo. DOI: https://doi.org/10.5281/ZENODO.7074521 (Accessed: 22 June 2024).

Ruth, N., Burghardt, M., & Liebl, B. (2023). From Clusters to Graphs. Toward a Scalable Viewing of News Videos. In A. Šeļa, F. Jannidis, & I. Romanowska (Eds.), *Proceedings of the Computational Humanities Research Conference 2023* (pp. 167–177). Paris: Computational Humanities Research. [= *CEUR Workshop Proceedings,* 3558] URL: https://ceur-ws.org/Vol-3558 (Accessed: 22 June 2024).

Ryan, M., & Lenos, M. (2020). *An Introduction to Film Analysis. Technique and Meaning in Narrative Film.* London: Bloomsbury Academic.

Salt, B. (1974). Statistical Style Analysis of Motion Pictures, *Film Quarterly,* 28(1), 13–22. DOI: https://doi.org/10.2307/1211438 (Accessed: 22 June 2024).

Id. (2006). *Moving Into Pictures. More on Film History, Style, and Analysis.* London: Starword Publishing. URL: http://www.starword.com/MovPicFin.pdf (Accessed: 22 June 2024).

Serengil, S.I., & Ozpinar, A. (2021). HyperExtended LightFace. A Facial Attribute Analysis Framework. In *Proceedings of the International Conference on Engineering and Emerging Technologies 2021* (pp. 1–4). Istanbul: IEEE Xplore. DOI: https://doi.org/10.1109/ICEET53442.2021.9659697 (Accessed: 22 June 2024).

Sikov, E. (2010). *Film Studies. An Introduction.* New York City: Columbia University Press.

Sittel, J. (2017). Digital Humanities in der Filmwissenschaft, *MEDIENwissenschaft. Rezensionen. Reviews,* 34(4), 472–489.

Souček, T., & Lokoč, J. (2020). TransNet V2. An effective deep network architecture for fast shot transition detection, *arXiv.* DOI: https://doi.org/10.48550/arXiv.2008.04838 (Accessed: 22 June 2024).

Springstein, M., Stamatakis, M., Plank, M., Sittel, J., Mauer, R., Bulgakowa, O., Ewerth, R., & Müller-Budack, E. (2023). TIB AV-Analytics. Eine webbasierte Plattform für wissenschaftliche Videoanalyse und Filmstudien. In H.-H. Chen & W.-J. Duh (Eds.), *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval 2023* (pp. 3195–3199). New York: Association for Computing Machinery Special Interest Group on Information Retrieval. DOI: https://doi.org/10.1145/3539618.3591820 (Accessed: 22 June 2024).

Stam, R. (2000). *Film Theory. An Introduction.* Malden, Mass.: Blackwell Publishing Limited.

Theiner, J., Müller-Budack, E., & Ewerth, R. (2022). Interpretable Semantic Photo Geolocation. In *Proceedings of the EEE/CVF Winter Conference on Applications of Computer Vision 2022* (pp. 750–760). Waikoloa: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.48550/arXiv.2104.14995 (Accessed: 22 June 2024).

Tseng, Ch., Liebl, B., Burghardt, M., & Bateman, J. (2023). FakeNarratives. First Forays in Understanding Narratives of Disinformation in Public and Alternative News Videos. In P. Trilcke, A. Busch, & P. Helling (Eds.), *DHd 2023. Open Humanities Open Culture.* Trier/Luxemburg: Verband Digital Humanities im deutschsprachigen Raum. DOI: https://doi.org/10.5281/zenodo.7715277 (Accessed: 22 June 2024).

Tsivian, Y. (2009). Cinemetrics. Part of the Humanities' Cyberinfrastructure. In M. Ross, M. Grauer & B. Freisleben (Eds.), *Digital Tools in Media Studies. Analysis and Research. An Overview* (pp. 93–100). Bielefeld: Transcript.

Vonderau, P. (2017). Quantitative Werkzeuge. In Hagener, M., & Pantenburg, V. (Eds.), *Handbuch Filmanalyse.* Wiesbaden: Springer VS [= *Springer Reference Geisteswissenschaften*]. DOI: https://doi.org/10.1007/978-3-658-13352-8_28-1 (Accessed: 22 June 2024).

Walkowski, N.-O., & Pause, J. (2018). Everything is Illuminated. Zur numerischen Analyse von Farbigkeit in Filmen, *Zeitschrift für digitale Geisteswissenschaften,* no pag. Wolffenbüttel: Herzog August Bibliothek. DOI: https://doi.org/10.17175/2018_003 (Accessed: 22 June 2024).

Weitin, T. (2017). Skalierbares Lesen, *Zeitschrift für Literaturwissenschaft und Linguistik,* 47, 1–6.

Wevers, M., & Smits, T. (2020). The visual digital turn. Using neural networks to study historical images, *Digital Scholarship in the Humanities,* 35(1), 194–207. DOI: https://doi.org/10.1093/llc/fqy085 (Accessed: 22 June 2024).

Wickham, H. (2016). *ggplot2. Elegant Graphics for Data Analysis.* Berlin/Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-319-24277-4 (Accessed: 22 June 2024).

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN. A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation 2006* (pp. 1556–1559). Genoa: ELRA Language Resources Association. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf (Accessed: 22 June 2024).

Wu, H., Chen, K., Liu, H., Zhuge, M., Li, B., Qiao, R., Shu, X., Gan, B., Xu, L., Ren, B., Xu, M., Zhang, W., Ramachandra, R., Lin, Ch.-W., & Ghanem, B. (2023). NewsNet. A Novel Dataset for Hierarchical Temporal Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023* (pp. 10669–10680). Vancouver: Institute of Electrical and Electronics Engineers. DOI: https://doi.org/10.1109/CVPR52729.2023.01028 (Accessed: 22 June 2024).

Wu, H.-Y., Palù, F., Ranon, R., & Christie, M. (2018). Thinking Like a Director. Film Editing Patterns for Virtual Cinematographic Storytelling, *ACM Transactions* *on Multimedia Computing, Communications, and Applications,* 14(4), 1–22. DOI: https://doi.org/10.1145/3241057 (Accessed: 22 June 2024).

Wu, L., Petroni, F., Josifoski, M., Riedel, S., & Zettlemoyer, L. (2020). Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing 2020* (pp. 6397–6407). arXiv: Association for Computational Lingutistics. DOI: https://doi.org/10.48550/arXiv.1911.03814 (Accessed: 22 June 2024).

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2022). Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *ICASSP 2023. IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1–5). Rhodes Island: IEEE. DOI: https://doi.org/10.1109/ICASSP49357.2023.10095969 (Accessed: 22 June 2024).

Zacks, J.M. (2010). Wie wir unsere Erfahrungen zu Ereignissen organisieren, *Psychological Science Agenda,* 24(4).

Zhang, H., Yuan, T., Chen, J., Li, X., Zheng, R., Huang, Y., Chen, X., Gong, E., Chen, Z., Hu, X., Yu, D., Ma, Y., & Huang, L. (2022). PaddleSpeech. An Easy-to-Use All-in-One Speech Toolkit. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies. System Demonstrations* (pp. 114–123). Seattle: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2022.naacl-demo.12 (Accessed: 22 June 2024).

Zhang, H., Li, X., & Bing, L. (2023). Video-LLaMA. An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. System Demonstrations* (pp. 543–553). Singapur: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2023.emnlp-demo.49 (Accessed: 22 June 2024).

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places. A 10 Million Image Database for Scene Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 40(6), 1452–1464. DOI: https://doi.org/10.1109/TPAMI.2017.2723009 (Accessed: 22 June 2024).

Figure Credits

Fig. 1–4 are self-created screenshots from the authors' work with TIB AV-A (Fig. 1) and the R package *ggplot* (Fig. 2–4). They are all published here for the first time.

IV. FORMS OF DIGITAL TEXT ANALYSIS

Python or *R*? Getting Started with Programming for Humanists

William Mattingly

Abstract This article describes how programming languages such as *Python* and *R* open up new research opportunities for the humanities by analysing large text corpora, visualising patterns in data and automating repetitive tasks. *Python* and *R* are probably the most prominent programming languages for engaging in the Digital Humanities. *Python* offers advantages for text analysis and machine learning due to its simple syntax and versatile libraries, while *R* scores with its statistical functions and visual representation options for data manipulation. The choice between *Python* and *R* therefore depends on the specific research requirements, although both languages are well suited to the humanities due to their strong communities and extensive resources. Learning strategies for getting started with programming and how to deal with potential pitfalls are also discussed.

Keywords Programming Languages, Python, R, Humanities, Digitization, Digital Education

1. Introduction

Programming is a fundamental skill in the hard sciences and maths. In these disciplines, programming languages like *Python* and *R* function as a tool to conduct research. They enable scientists to process large datasets, run complex simulations, and automate repetitive tasks. For instance, a biologist might use *Python* to analyze billions of genetic sequences from many organisms, or a statistician might use *R* to model and visualize patterns across data. In this chapter, we will explore how this concept translates to humanities data.

2. Data and the Humanities

Unlike the sciences, the humanities are traditionally associated with qualitative analysis, valuing a close reading of source material. For most of the twentieth century the majority of the humanist's data was available in analog forms, that is through the medium of print. The methods we developed for analyzing our data, therefore, were targeted at the practical limitations of our data and human anatomy. Even the

William Mattingly: *Python* or *R*? Getting Started with Programming for Humanists, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 177–187. DOI: https://doi.org/10.11588/heibooks.1521.c21947

most ardent researcher cannot possibly read the entire corpus of Latin literature in a single lifetime. Presuming one had the capacity to read such a quantity of material, one would have to physically obtain it. In manuscripts, these resources are scattered across different continents inside institutions that have limitations to access. Even presuming one could overcome these physical limitations, one would then have to synthesize it in some meaningful way. How can we retain all this information and translate that knowledge to an audience practically? We simply can't.

During the second half of the twentieth century, the medium for transmitting data changed. Text could be rendered as numerical data. Initially this was via punched cards, such as those used by *Index Thomisticus,* the first project that digitally rendered the collected works of Thomas Aquinas (approximately 10 million words, Busa 1980; cf. the chapter from J. Peters in this volume, p. 316). Because the digital data was rendered on physical punched cards, it still required physical access to the punched cards. These could, however, be transmitted differently than print and loaded on different machines across the globe.

As technology continued to develop, data could be stored not as physical punched cards, but as disks. As these disks became smaller and capable of holding large quantities of data, these disks could be transported more easily and reproduced more cheaply. By the 1990s, the *Index Thomisticus* evolved with this technology and made its data available via CDs (Economist 2020). This meant that scholars could purchase identical data and analyze it regardless of location. This data could be reproduced, distributed, and computationally accessed by a larger audience.

Over the last two decades, we have not only produced larger quantities of digital data, accessibility likewise increased thanks to the advent of the internet and the cloud. The cloud allowed for data to sit on a server in one place and be accessed by someone on the other side of the globe. We can, for example, study the entire *Patrologia Latina* from a beach in Florida via the open-access *Corpus Corporum* project at the University of Zurich.[1] So ubiquitous is this technology today, that it is sometimes difficult to appreciate how incredible this feat truly is.

Today, the limitations of quantity and access are gradually vanishing. As the scale of humanities research extends into the realm of *big data,* the ability to handle and interpret this information becomes crucial. What do we do with this data? How do we access it systematically? How do we use it to frame questions and translate that knowledge into something useful? As we will see throughout this chapter, programming affords us potential solutions to these questions. Programming is also an emerging skill that departments are including in their curricula, such as that of the recent Ph. D. in Digital History at Clemson University.

With the help of programming languages like *Python* and *R,* humanists can analyze large corpora of texts, visualize complex relationships in data, and uncover

1 See https://mlat.uzh.ch (Accessed: 22 June 2024).

patterns that would be impossible to find manually. Whether it's studying the entire collected works of Augustine (d. 430) or mapping the social network of Alcuin (d. 804), programming opens up a whole new world of possibilities for humanistic inquiry. For the humanities, much like the maths and sciences, programming languages are not necessarily used for creating software; scripting languages like *Python* and *R* function like tools by which we can analyze large data, interpret it, and visualize findings. Unlike software, scripting allows for researchers to leverage the quick utility of programming without long-term sustainability necessary to maintain software.

3. Why Learn to Code?

One of the most important questions we can ask is *why learn to code?* Behind this question lie two other questions: *What is the utility of coding and how does it benefit me directly?* To answer these questions, let us consider a problem. Imagine we needed to identify all named persons in the writings of Augustine. We could, of course, spend months or years going through and manually tagging each person. This would be time consuming and repetitive. If we knew programming, however, this problem becomes far simpler to solve and the solution can be developed in hours (or days depending on its complexity). We could either construct a set of rules for all the people we expect to find in the letters or (if we did not know this), we could develop a machine learning solution that would learn the features of the words that correspond to people and classify them for us automatically. This task is known as *named entity recognition* (cf. the chapter from E. Gius in this volume).

This process by which we create a solution that can be implemented repetitively over similar data, is known as automation. Automation is one of the key reasons to learn to program. As humans, we are very bad at doing the same task repetitively and consistently. Computers, on the other hand, are perfect at both of these tasks. Learning to program is, in part, learning to automate tasks and is one of the key benefits. It allows the researcher to spend more time researching and less time performing repetitive tasks.

In addition to automation, programming also affords us the ability to develop solutions that work at scale; this means being able to perform practically the same solution across millions of data. If we can identify all persons in a text of 10,000 words in five seconds (a reasonable time), then we could easily repeat this process across millions of texts in just a few hours, depending on computational resources.

Through programming, we can also develop and apply machine learning solutions with a few lines of code. Continuing with the example above, imagine we needed to solve the same problem with the works of Jerome. Here, we could take the machine learning solution we developed for the works of Augustine and apply it to the works of Jerome and have comparable results.

Programming allows humanists to achieve far more than the creation of novel solutions. It fundamentally changes the way we frame problems. A knowledge of programming brings with it knowledge of solutions or potential solutions to unknown problems (or unasked questions). Those who have a knowledge of programming and know what is possible to do with it know that they can ask new questions and pursue novel research directions. It makes it feasible to conduct studies at a scale and depth previously unimaginable. With programming, researchers can identify patterns and trends across large datasets, leading to new insights and understandings.

Selecting the right programming language largely depends on the problem at hand. If the researcher needs to build a website with custom functionality, then learning HTML, *JavaScript,* and *React* (a way of building JavaScript components easily) may make the most sense. If the researcher needs to work with and manipulate data in any way, there are two dominant languages that should be considered: *Python* and *R.*

4. The Benefits of *Python* and *R*

Python and *R* have emerged as the primary programming languages for many fields, namely data science, machine learning, natural language processing, statistics, the sciences, the maths, and the humanities. *Python* was created by Guido van Rossum in the 1980s with the idea that code should be easy-to-read and easy-to-right with a syntax, or style of writing, that was concise and simple (Van Rossum & Drake 1995). Against *Python,* sits *R* which was created by Ross Ihaka and Robert Gentleman in the 1990s (R Core Team 2021). Unlike *Python, R* was designed purely for statistical analysis. Its syntax is a bit unconventional, but it affords users the ability to apply statistical methods to quantitative data easily and visualize the results.

Both of these languages have increased in popularity largely due to their excellent and active communities. These communities write packages, or libraries, for each programming language. A package is a collection of classes and functions (think of these as large blocks of code) that can be leveraged by others in the community. This means that new users to the programming language can do complex tasks with very little code, making them ideal programming languages for beginners. For example, students new to *Python* can load a machine learning model and perform *named entity recognition* on the entire corpus of Thomas Moore (for NER cf. the chapter from E. Gius in this volume). This is thanks to the natural language processing library *spaCy*[2] and the contributions to that project by Patrick J. Burns who created *LatinCy* (Honnibal & Montani 2017; Burns 2023). If one were interested in doing *transformer-based topic modeling,* which is the newer approach to this decades-old methodology, a researcher can do so with just 2 lines of code using *BERTopic* (cf. the

2 See https://spacy.io (Accessed: 22 June 2024).

chapter from M. Althage in this volume, esp. p. 259, n. 19). While further methodological steps are necessary to refine a researcher's approach, such as topic identification, validation, adjusting hyperparameters, the code required to do so remains minimal. This is because *BERTopic* employs many advanced methods in sequence for the user. It even allows for the quick visualization of a topic model. Contributions like these make *Python* and *R* appealing to those with limited and advanced coding experience alike. This appeal, in turn, fosters a healthy community which continues to grow. As the community grows, more members contribute their own packages. As time progresses, the cycle continues to repeat.

5. Comparing *Python* and *R* for Humanities Research

The relevance of *Python* and *R* to humanities research is multifold. *Python*'s simplicity and readability make it a great starting point for humanists new to programming. Its extensive libraries such as *Pandas*[3] for data manipulation, *spaCy* and the *Natural Language Toolkit* for natural language processing, and *Matplotlib*[4] or *Seaborn*[5] for visualization, provide valuable tools for various research tasks in the humanities. On the other hand, *R*'s strong data handling and built-in statistical capabilities, as well as its powerful visualization libraries, make it particularly suited for humanities researchers who work extensively with statistical data or need to create complex visualizations. When it comes to choosing between *Python* and *R* for humanities research, the decision often boils down to personal preference, specific project requirements, and the kind of data you'll be working with. Both *Python* and *R* have their strengths and are capable tools for handling humanities data.

Python has a few key advantages. First, its syntax corresponds to other programming languages and is generally easier for those new to programming. Second, *Python* allows users to quickly build websites via libraries like *Django*[6] and *Flask* (Grinberg 2018). Through *Streamlit*[7], a novice Python programmer can build a custom data-based application and put it in the cloud with only a few lines of code. Third, Python is often the first choice within the machine learning community. This means that most recent machine learning developments are available first in *Python.* Fourth, most natural language processing advances are often developed in *Python,* making it ideal for tasks like text classification, *topic modeling,* and *named entity recognition.*

3 See https://pandas.pydata.org (Accessed: 22 June 2024).
4 See https://matplotlib.org (Accessed: 22 June 2024).
5 See https://seaborn.pydata.org (Accessed: 22 June 2024).
6 See https://djangoproject.com (Accessed: 22 June 2024).
7 See https://streamlit.io (Accessed: 22 June 2024).

R, on the other hand, was built for statistics. It has several advantages over *Python.* First, *R*'s syntax and functionality are tailored for statistical modeling, allowing for complex analyses with concise code. Second, *R* provides an extensive collection of packages like *ggplot2* (Wickham 2016) and *Shiny*[8] that enable high-quality data visualization and interactive web applications. While *Python* boasts of good visualization libraries like *Plotly*[9] and *Seaborn* (Waskom et al. 2017), the visualizations in *R* are easier to produce, tend to look nicer, and are easier to customize. This means that users can not only analyze data but also create visually appealing representations with ease. Third, *R*'s integration with various data sources and its data manipulation capabilities through packages like *dplyr* make it a powerful tool for data wrangling. Fourth, although *R*'s machine learning capabilities may not be as extensive as *Python*'s, packages like *caret*[10] and *randomForest*[11] still provide robust tools for machine learning.

Both languages have active and supportive communities, so you'll find plenty of resources and help for both. Choosing between *Python* and *R* will depend on your specific needs and goals in humanities research. There are three main questions to consider. First, consider your own research needs. If your research involves a lot of text analysis or natural language processing, *Python* might be the better choice due to libraries like *spaCy.* If your work involves heavy statistical analysis or you need to create detailed visualizations, *R* might serve you better. Second, consider your own learning style and your experience with programming. If you're new to programming, *Python* may be easier to learn. It will be important to look at a couple snippets of code in *Python* and *R* to get a sense of how different these two languages are. Third, consider the communities behind both programming languages. Those communities will be there to help you when you run into challenges. Both *Python* and *R* have strong communities, but depending on your area of study, one might have more relevant resources and discussion boards than the other.

In the end, there's no definitive right or wrong choice between *Python* and *R* for humanities research. It's about choosing the tool that best suits your needs and complements your research. A humanist interested in using programming in their research will likely learn to write in both languages but typically prefer one over the other. This is because most things done in one of these languages can be done in the other, even though it may not be as easy. If you have a project entirely written in *Python,* for example, but need to produce a nice visualization, it may not make sense to introduce *R* into the workflow for one step. Instead, you will write more code and use the Python library *Seaborn.* Likewise, if you are presenting your findings on text

8 See https://shiny.posit.co (Accessed: 22 June 2024).

9 See https://plotly.com/python (Accessed: 22 June 2024).

10 See https://cran.r-project.org/web/packages/caret/vignettes/caret.html (Accessed: 22 June 2024).

11 See https://cran.r-project.org/web/packages/randomForest/randomForest.pdf (Accessed: 22 June 2024).

analysis and statistics, it may not make sense to use *Python* and *spaCy* for lemmatization, or the reduction of all words to their root form.

6. Getting Started with Programming

One of the most challenging aspects of learning to program is installing the programming language on your computer. Each operating system, such as Mac, Windows, or a Linux distribution (like Ubuntu), requires you to install the language differently. Each operating system also has unique steps. On Windows, for example, you need to make sure that *Python* sits in your system's PATH (often manually). Each operating system also introduces small, but critical differences. On some Macs, for example, you will have *Python* 2 pre-installed on your system. This means that when you install the recent version of *Python* (*Python* 3.12 as of writing this), you will have two versions of the same programming language on your computer. This means that you need to use the command "python3" in the command line to execute a *Python* file on some Macs. On Windows and Linux, however, "python" will be the command you use. Many of these issues are negated, however, if one uses virtual environments or *Conda*[12].

Installing the programming language is only one hurdle. Users usually want to also install a custom way to interact with the programming language. For *Python*, this typically means installing an integrated development environment (IDE) like *JupyterLab* or *VS Code.* For *R,* this means installing *R Studio.* These are tools that allow you to write and execute code in a single space. They allow you to learn more easily and also it will be the way in which you typically engage with a programming language. Installing and setting up an IDE will vary depending on your system requirements. Again, this is a step that can lead to confusion and issues as well.

In my experience, the frustrations that new students encounter during this process can dissuade them from wanting to learn to program. These are some of the most important moments in a student's career. It is when they are curious about programming and eager to learn. These frustrations can quickly dim the light of curiosity. To avoid this, I recommend that all new students skip the installation of the programming language and the IDE entirely. Instead, there are numerous companies that offer cloud-based solutions to these issues. They allow you to access a server remotely and run *Python* from your browser.

Constellate from ITHAKA[13] is one such solution, but it requires institutional access. It provides students with a virtual environment with enough resources to even do machine learning should they wish. It comes pre-installed with libraries commonly needed on humanities projects. Each user instance also comes pre-installed with

12 See https://docs.conda.io/en/latest (Accessed: 22 June 2024).
13 See https://constellate.org (Accessed: 22 June 2024).

JupyterLab, an IDE that facilitates data management and learning (via *Jupyter notebooks*). This makes it ideal for classroom settings. I have taught for three years with *Constellate* and highly recommend it.

Not all students will have institutional access to *Constellate,* however. If this is your case, there are comparable services available. The most popular is *Google Colab*[14] which has several tiers, including a free version. It can link to your *Google Drive.* This means you can load data onto your *Google Drive,* interact with it, manipulate it, and save it. The free tier tends to drop occasionally, but works well for getting started nonetheless.

7. Resources for Learning

Learning to program as a humanist is often an individual endeavor. Even if you have a formal education in *Python* in a college classroom setting, you will have to rely on your ability to self-teach for the remainder of your programming career. This is because real-world data and problems are messy. Clear and easy solutions are rarely evident. You must be prepared to learn new aspects of a programming language to solve novel problems as they emerge. Imagine you have learned to program in *Python* to do *named entity recognition.* Now, you need a way to visualize those named entities in a network graph. How do you do that? In this case, it would be time to learn *NetworkX*[15] to collect the data and either *Matplotlib* or *PyVis*[16] to visualize it.

Because *R* and *Python* have large communities and a lot of libraries available for solving common problems, there are numerous resources available for furthering your education. One great resource is Walsh (2021), an open-source textbook titled *Introduction to Cultural Analytics and Python.* This textbook not only teaches the basics of *Python,* it also provides an introduction to a few key methodologies, such as text analysis and network analysis. The standard textbook for Humanities data and *R* remains *Humanities Data in R* by Arnold & Tilton (2015). It has open-source online resources, but these are meant to be used alongside the purchased textbook. Another *R*-specific resource is the open-source textbook *Computational Historical Thinking. With Applications in R* by Mullen (2018). Unlike recent open-source textbooks (such as Walsh's), this is not designed with *JupyterBook.* Nevertheless, it still has useful supplementary material, such as worksheets.

After you have acquired a basic understanding of either *Python* or *R,* you will find yourself frequently needing to learn a specific library. Often, the documentation for these libraries is sparse. It is usually written by experts for more advanced users.

14 See https://colab.google (Accessed: 22 June 2024).
15 See https://networkx.org (Accessed: 22 June 2024).
16 See https://pyvis.readthedocs.io/en/latest (Accessed: 22 June 2024).

At these times, it can be necessary to have a relatable tutorial. If you want to stick with academic tutorials, the *Programming Historian*[17] is probably the best resource available. As of now, there are 101 lessons for both *Python* and *R.* These often center around a specific problem or methodology. The lessons published here are both open-source and peer reviewed via *GitHub.* They tend to be more targeted at specific problems or specific libraries.

Some of the best resources available, however, are written by non-academics. These resources are published on *YouTube* and *Medium.* The easiest way to find resources for your specific problem is to search on these platforms for something in which you are interested.

8. Common Pitfalls with Programming

As you progress in your programming career, you will encounter many pitfalls. One of the most common is a bug in your code. Most programming languages will give you an error message to indicate why a piece of your code failed and it will often point you to the specific line. As you work with external libraries and write more complex code, however, debugging can be trickier. Fortunately, there is (or was) a healthy community at *Stack Overflow* which allows users to post bugs and ask for help. Usually someone in the community will respond within a few hours. Since the advent of *ChatGPT,* however, traffic on *Stack Overflow* has decreased. The implication is that more users are asking questions via *ChatGPT* or some other similar service. For basic coding issues, *ChatGPT* will provide fairly good and specific advice for fixing a bug. There is a good reason for this; it was trained on a lot of *Stack Overflow* data. When using *ChatGPT* or *Stack Overflow,* it is always important to not simply copy-and-paste the solution, but to understand *why* the bug surfaced in the first place so that you can correct the issue *and* learn from your mistakes.

Another common pitfall is the use of algorithms that are not fully understood by the programmer. While a programmer does not need to know how neural networks work to leverage them and generate useful output, it would be unwise to rest an argument on the statistical output of a model whose algorithm the researcher does not understand. When you do not yet understand the methods fully, therefore, programming should be used as a tool to assist in research, never as a tool to validate arguments.

17 See https://programminghistorian.org (Accessed: 22 June 2024).

9. Conclusion

Programming is not meant to replace traditional humanistic inquiry; instead, it provides humanists with new avenues to conduct research. It allows us to frame questions that we could not otherwise answer. It allows us to automate tasks in hours that could otherwise take years. And it allows us to glean new insights from large quantities of data. As the humanities evolves and technology becomes ubiquitous, tomorrow's humanists will likely acquire more technical skills, just as they did with the advent of the Word processor. Tomorrow's Word process is programming.

References

Arnold, T., & Tilton, L. (2015). *Humanities Data in R. Exploring Networks, Geospatial Data, Images, and Text.* Cham: Springer. DOI: https://doi.org/10.1007/978-3-319-20702-5 (Accessed: 22 June 2024).

Burns, P.J. (2023). LatinCy. Synthetic Trained Pipelines for Latin NLP. *arXiv.* DOI: https://doi.org/10.48550/arXiv.2305.04365 (Accessed: 22 June 2024).

Busa, R. (1980). The Annals of Humanities Computing. The Index Thomisticus, *Computers and the Humanities,* 14(2), 83–90. URL: https://www.jstor.org/stable/30207304 (Accessed: 22 June 2024).

The Economist (2020). How data analysis can enrich the liberal arts. URL: https://www.economist.com/christmas-specials/2020/12/19/how-data-analysis-can-enrich-the-liberal-arts (Accessed: 22 June 2024).

Grinberg, M. (2018). *Flask web development. Developing web applications with python.* Sebastopol: O'Reilly Media.

Mullen, L.A. (2018–). Computational Historical Thinking. With Applications in R. In *Computational Historical Thinking [Blog/Preprint].* URL: https://dh-r.lincolnmullen.com (Accessed: 22 June 2024).

R Core Team (2021). R. A language and environment for statistical computing [Software]. Wien: R Foundation for Statistical Computing. URL: https://www.R-project.org (Accessed: 22 June 2024).

Van Rossum, G., & Drake Jr, F.L. (1995). Python reference manual. Version 1.2. Amsterdam: Centrum voor Wiskunde en Informatica. URL: https://ir.cwi.nl/pub/5008/05008D.pdf (Accessed: 22 June 2024).

Walsh, M. (2021). Introduction to Cultural Analytics & Python. Version 1. Zenodo. DOI: https://doi.org/10.5281/zenodo.4411250 (Accessed: 22 June 2024).

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D.C., Augspurger, T., Halchenko, Y., Cole, J.B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, M., Ram, Y., Yarkoni, T., Williams, M.L., Evans, C., Fitzgerald, C., Fonnesbeck, B. Ch., Lee, A., & Qalieh, A. (2017). mwaskom/seaborn. Version 0.8.1 [Python-Package]. Zenodo. DOI: https://doi.org/10.5281/zenodo.883859 (Accessed: 22 June 2024).

Wickham, H. (2016). ggplot2. Elegant Graphics for Data Analysis. Cham: Springer [= *Use R!*]. DOI: https://doi.org/10.1007/978-3-319-24277-4 (Accessed: 22 June 2024).

Stylistic Analysis

Fotis Jannidis

https://orcid.org/0000-0001-6944-6113

Abstract In the humanities, the term *style* usually refers to a systematic choice of means of expression in a sign system, e.g. language, that is characteristic of an author or a genre or an epoch, etc. Stylometry uses these features, in the case of texts, e.g. lexis, syntax, semantics, and text structure, to attribute authorship, to profile authors, or even to assign periods and genres by means of quantitative methods such as clustering or classification. Stilometric methods were applied to religious texts very early on and, as can be seen from the history of the analysis of the Pauline epistles, reflect important stages in the development of stilometry from univariate to multivariate analysis, which today is usually carried out in a probabilistic framework with numerous test repetitions.*

Keywords Stylometry, Authorship, Genre

In many humanities disciplines, the term *style* is used to describe a use of language and other means of expression determined by an author's conscious or subconscious choice and not dictated by the content alone. Typical uses can be distinguished: style is understood as an aspect of form and refers to the entire text or even to an author's oeuvre. Style is often understood as an expression of the author's individuality, which can be seen in the selection of individual elements specific to an author. This selection might deviate systematically and characteristically from the prevailing conventions (Hermann et al. 2015). Stylistic analysis is then usually understood, especially in the study of literary texts, to indicate the identification of linguistic devices typical of the respective style, i.e., specific sentence constructions, verb constructions, adjective sequences, etc. (Leech & Short 2007).

1. Stylometry

Stylometry is the corpus-based analysis of style using quantitative methods. Compared to a qualitative description of the characteristics of a style, the aim of stylometry is much simpler: to assign a text to a group of other texts based on stylistic characteristics. Stylometry is one of the most productive fields in the Digital Humanities with a

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

Fotis Jannidis: Stylistic Analysis, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 189–202. DOI: https://doi.org/10.11588/heibooks.1521.c21948

long history (Holmes 1998) and many different approaches and applications in many academic disciplines and fields of application. Research overviews with large bibliographies can be found in Juola (2008), Stamatatos (2009), and Neal (2017).

The focus is primarily on tasks dealing with aspects of authorship:

- Authorship attribution: An anonymous text is assigned to an author from a set of possible candidates. This procedure can only work if texts exists where the authorship of the texts is undoubted.
- Authorship verification: Two texts are given; the task is to determine whether they were written by the same author.
- Author profiling, also known as *sociolinguistic profiling* (Grant 2022, 20): The task is to determine the gender, age, native language, personality traits, illnesses such as dementia (Hirst & Wei Feng 2012), or profession of the author based on one or more texts.
- Multi-authorship: In texts written by several authors, the sections should be assigned to their respective authors.
- Stylochronometry: The task is to determine the chronological order of a series of texts based on stylistic characteristics, such as determining the chronology of an author's texts (Seminck et al. 2022).

In addition to investigating these aspects of authorship, stylometry also addresses several other aspects, such as the attribution of a text to a period or genre based on its stylistic characteristics (Jannidis & Lauer 2014), the recognition of the physical writing situation (dictation, handwriting, typewriting, etc., Hoover 2021), the visibility of translators (Rybicki 2021), and character idiolects (Burrows 1980). Stylometry is also used to detect plagiarism and to reorganize texts so that their style no longer corresponds to that of the author. There is also a relatively independent sub-field of stylometry within forensic linguistics, which deals for example with the classification of blackmail letters. Forensic stylistics also uses corpora to identify author-specific usage, such as a deviation in style. However, given the limited amount of text in most cases and the parameters of the legal system, stylistics has developed its own practices (Grant 2022).

Authorship attribution, which is one of the most frequently processed tasks, takes place in very different contexts. In the simplest case, it involves assigning an unknown longer text to one of two authors from whom numerous longer texts of the same genre and from the same period are available. However, variations of these factors quickly lead to much more complex research groupings. For example, there could be a high number of possible authors. The question of whether the author of the text in question is certainly among the candidates *(closed set)* or not *(open set)* is decisive for the methods. In addition, often only texts from other genres, other time periods, or in different languages are available. At times, there are only short texts available, which complicates the statistical analysis.

The quantitative analysis of style uses text features to group texts according to authorship *(cluster analysis)* or to classify them *(supervised learning).* These features can be divided into four groups (Neal 2017, 11):

- Lexical, such as word frequency, word n-grams, number of letters, or average word length
- Syntax, such as frequency of punctuation marks, sentence length, complexity of sentence structure
- Semantics, such as synonyms and topic distribution
- Text Structure, such as length of paragraphs, paragraph indentations, and font

While the step in the quantitative analysis of style concentrated on rare words and sentence constructions as individualizing features, it has long been common practice to consider the distributions in the use of the *most frequent* features as particularly discriminative. Research has shown that there is no clearly defined set of features that leads to the best results in each of the tasks and in each of the many possible interrelations. In the beginnings of stylometry, a single feature was collected and compared; however, since the 1980s, the more common practice has been to use several features at once.

The *Shared Tasks* for stylometry from PAN[1] have been available since 2011. *Shared Tasks* are an established method in computational linguistics for summarizing and further developing research knowledge on a problem. For this purpose, a data set is published together with a task that is to be solved using the data. For example, one of the tasks in 2019 was to identify authors of fanfiction from a specific fandom; however, only their works on other fandoms were known *(cross-domain),* and the task was formulated as an *open-set attribution.* The submitted solutions are then evaluated using data that the participants have not yet seen, and the result and the data sets are published so that later studies can compare their results directly with those of the competition. These PAN *Shared Tasks* have furthered standardization and promoted knowledge about the methodological state of stylometry. They remain cutting edge in the field.

In the context of stylometry, the specific linguistic form of a text is the result of several factors. The factors include: The theme of the text, the general language usage at the time, which is often abbreviated as period style, the linguistic features typical of the text type (genre style), and the author-specific linguistic choices (authorial style). The aim of a stylometric study is to make a classification based on linguistic characteristics that do not essentially depend on the topic, thus considering authorship attribution, and to control other factors like epoch and genre. The factors authorship, genre, and epoch/period have a simultaneous influence on the style; accordingly,

1 See https://pan.webis.de (Accessed: 16 June 2024).

research designs must control the respective non-questionable aspects as possible confounding variables. In the case of authorship attribution, one can control by including texts from the same period and genre in the corpus.

After compiling all relevant texts in digital form, the N features deemed relevant, e.g., words, punctuation marks, letter n-grams etc., are extracted from all texts, where N stands for the number of features. Each text is then represented as a vector with the N values of these features. Geometrically, each of these vectors can be seen as a point in an N-dimensional space. Distance measures then allow one to determine the distance between these points, each of which represents a text. Let's assume the simplest constellation. There is a text X with unknown authorship and two groups of texts, G_A and G_B, where the first is by author A and the second by author B, and we know that one of the two is the author of X. We now select the text features so that the texts have a relatively small distance within the group and a clear distance between the groups. If we now add the text X in question, we will likely observe that the text has a relatively small distance to one group and a relatively large distance to the other – if the above-mentioned conditions are met, that the texts are of the same genre and from the same period, and that the text in question really is by one of the two authors.

However, a decision in favor of a certain set of features that optimally differentiates the texts of the candidates can only be made if it is certain that only one of the candidates can be the author *(closed set).* In most cases, however, there is no certainty; it could also be another author who is not among the candidates *(open set).* In this case, however, the text features cannot be optimally selected. From a practical research perspective, the procedure of selecting features that are considered optimal for a certain data set and only one evaluation method has also not been successful as the results are hardly verifiable and comparable. For this reason, newer approaches are based not only on working with a specific set of characteristics and then carrying out a distance measurement; rather, the measurement is repeated with a different selection of features (Juloa 2015).

In their methodological study on authorship attribution under *open set* conditions, Sperling et al. have shown that Stephen King can be clearly identified as the author of the novels published under the name Richard Bachmann (Sperling et al. 2023). To do this, they collected a large number of characteristics (120,000) based on the standard imposter method (Koppel & Winter 2014) and repeatedly drew a smaller sample (10,000). They then use these 10,000 sample features to measure the distance between an excerpt of a novel by Bachmann and randomly selected excerpts from novels by four authors of horror novels including King. This process is repeated thousands of times for each excerpt of a novel by Bachmann. Each of these measurements results in different distances between the candidate authors and Bachmann, which can be noted as a ranking. It turns out that King ranks first more often than the other authors, i.e., is closest to the Bachmann text.

In analyses of this kind, as is already known from the long tradition of *Information Retrieval,* the determination of the distance measure, with which the distances between the vectors are measured, is decisive. A particularly important measure of this type is *Burrows Delta* (Burrows 2002), which has proven to be particularly robust (Evert et al. 2017). In the case of very large amounts of text, supervised machine learning methods have also proven to be very good (Savoy 2020).

The assumption that every author, in fact every person, when using language, has a unique way of using language, or in linguistic terminology an idiolect, is one of the basic theses of stylometry (Nini 2023). However, this assumption is not necessary to justify the way stylometry functions. It is sufficient to assume that there are discriminative features for a given set of authors. The first point is a difficult assumption to prove; the second point has now been empirically demonstrated for many constellations, but no claim to general validity can be derived. Overall, the findings of stylometry, namely that an author's texts exhibit demonstrable commonalities, are a problem for theories that model individuality as a mere discourse phenomenon or as an emergent phenomenon of interactions (Jannidis 2014, Grand & McLeod 2018). However, the metaphor of the *linguistic fingerprint* sometimes used in popular contexts is misleading. On the one hand, the idiolect develops over time, however slowly, and on the other hand, a corpus-specific and task-specific selection of linguistic features is used for the processing of stylometric tasks and not a constant identity marker. Overall, stylometry is still largely a science driven by individual studies, but in the last ten years it has increasingly been striving for a theoretical and methodological foundation (cf. Juola 2015, Nini 2023).

2. Using Stylistic Analysis for Religious Texts

In 1851, the mathematician Augustus de Morgan, who is now regarded as one of the founders of formal logic, speculated on the possibility of distinguishing the letter of Hebrews from the letters of Paul based on the average word length of the letters (De Morgan 1882, 215f.). This hypothesis is still regarded as one of the origins of stylometry (Holmes 1998, 112). Stylometric methods were then applied to biblical texts at an early stage, which showed a promising trajectory for the methods from the outset. As Radday wrote in a 1973 study, "certain problems that have vexed Biblical scholars for centuries can at least be approached and possibly solved once and for all by quantification" (Radday 1973, 273). Particularly in the case of controversies about authorship, stylometric methods are a good way of integrating a further source of information into the judgment process alongside historical and linguistic information (Oakes 2014). A brief overview of the question of the authorship of the Pauline letters shows how stylometric methods have been used and how the methods have changed over time.

The question of whether the fourteen Pauline letters were really written by the historical Paul or whether some were written by one or more other authors, whether the two letters to the Corinthians, Galatians, and Romans were actually written by one author, is one of the oldest topics of theological stylometry. Authorship questions have been discussed since the 18th century, not least because theologians perceived clear stylistic differences. Some of the research on this topic has also used simple quantitative arguments relating to the frequency of a word when comparing texts. Robert Morgenthaler's work has promulgated this comparison, listing the frequencies of all the words in the books of the New Testament (Morgenthaler 1958). One example can be found in Bujard's work on Colossians. Bujard only uses the simple counts (often based on Morgenthaler) and does not carry out a statistical test to see whether the observed differences are significant (Bujard 1973). The theologian Andrew Q. Morton broke with this established way of working in two ways: he used the computer, and he used statistical methods to support the reliability of his comparisons.[2] Morton's 1963 newspaper article was controversial because he claimed, based on his stylometric studies, that only four letters were written by the same author (1. and 2. Corinthians, Galatians, and Romans), whereas the historical scholars of his time assumed that the first letter to the Thessalonians, Philippians, and Philemon were also written by the historical Paul. The fierce criticism prompted him to present his method and results in detail (Morton & McLeman 1966).

Morton's work has yet to be fully appreciated, as the view of his achievement is overshadowed by his later work. His *Cumulative Sum* (CUSUM) technique was supposed to show clear authorship attribution even of short texts and was also used in court in Great Britain in the 1990s. However, the public success was countered by the doubts of specialist colleagues, who considered the statistical procedure and its application by Morton to be highly unreliable, and with good reason (Holmes 1998, 114). This perspective still distorts the reception of Morton's early work.

In his 1966 study, Morton essentially used two text features: the average sentence length and frequent words, so-called function words. He excludes the articles, which occur most frequently, since they depend too much on the subject of the text, and concentrates on *kai* (and), the particles *de, en* (in), *einai* (to be), and *autos.* First, he uses a corpus of ancient Greek texts to determine whether these function words are suitable for distinguishing between authors of a text by dividing the texts into segments of equal size, recording the frequency of the respective function word, and then using Chi^2 to calculate the probability of whether the observed difference between the segments is coincidental – in the latter case, this would speak in favor of two authors. All tests confirm the result: 1. Corinthians, 2. Corinthians, Galatians and Romans were written by one author. Philemon is problematic given its brevity, but the data also speak in favor of it belonging to this group. Morton's results confirm theses that were already formulated in the 19th century by F.C. Baur (Neumann 1990, 2ff.). A major

2 Neumann lists stylometric works that were not considered, yet worked with statistical analysis even before Morton (Neumann 1990, Chap. II).

weakness of his approach is the fact that each study only includes one variable, such as the frequency of *kai*, whereas multivariate analyses of the kind described above have become the standard since the 1980s at the latest.

Neumann, whose work was not published until 1990, yet was probably completed in 1980, collects 617 features, including lexical, syntactic, and other features. He carries out a detailed preliminary investigation into the extent to which these features divide a text corpus 100 % correctly in a discriminant analysis and then selects only the four features (word length, indefinite pronouns, words beginning with the letter *tau*, and the position of the first noun in the sentence) for his study of the Pauline epistles. His conclusion is that the disputed letters – Colossians, Ephesians, 2. Thessalonians – are attributed to Paul. The requirement that the features should divide the data 100 % correctly would currently be labeled as overfitting, as the feature selection for the textual data set fit exactly, yet it remains unclear how well they fit the actual object of study.

The philosopher Antony Kenny, who wrote an introduction to stylometry and its statistics in 1982 (Kenny 1982), published his relevant work on the New Testament four years later (Kenny 1986), in which he also deals with the problem of Paul's letters. He selects most of them from a set of 99 features that he developed for his work; these include conjunctions, particles, prepositions, the cases of nouns, pronouns, and adjectives to differentiate the texts. Kenny analyzes each of these features individually and then tests whether the frequencies differ significantly in the texts in question. His results are presented with caution. He concludes that some of the features in some letters seem to indicate a different authorship: the variance observed in twelve of the letters could also be explained by the fact that they are the work of a single, extremely versatile author. In his review of Kenny, Mealand claims that the result is not insignificantly dependent on his task, namely whether a text is particularly divergent, and shows that the correlations Kenny lists can also be evaluated so that the close relationship between Romans, 1. Corinthians, 2. Corinthians, and Galatians becomes clear (Mealand 1988).

The studies in the 1990s use multivariate statistical methods. Greenwood groups the texts based on the 10, 20, and 30 most frequent words and finds that the clusters are very similar (Greenwood 1993). He also uses a method that not only forms clusters, but also allows him to determine whether the clusters overlap in high-dimensional space, which would indicate that they are not separable units. He draws this conclusion even if the structure is not preserved in the PCA, a dimension reduction method that represents the high-dimensional data in two dimensions. In other words, there is an information structure that enables a clear separation of the texts, but it is only visible in high-dimensional space. Overall, however, his results show a strengthening of Morton's thesis, revealing that the function words make authorship attribution possible, a thesis "not comfortably assimilated within the spirit of classical scholarship" (217), despite being established in literary studies at the same time by the work of Burrows.

Ledger (1995) divides the texts into 1000-word sections and for each section collects 29 features that have previously proved useful in Plato's analysis, 19 of which are the proportion of words with a particular letter, 9 with the proportion of words with a particular final letter, and the standard measure of lexical complexity, type-token ratio. He also uses PCA and examines the resulting groupings. He recognizes a central cluster with 1. Corinthians, 2. Corinthians, Galatians, Philippians, 2. Thessalonians, and Romans as well as a second group with Colossians, Ephesians, and Hebrews; the others cannot be clearly assigned. His attempts to identify the characteristics that lead to the special position of Hebrews and to relate these in turn to the texts are noteworthy.

Like Neumann, Mealand uses a discriminant and factor analysis to reduce dimensions in his study, which also examines other NT letters in addition to the Pauline letters. This is based on 25 features, including the most frequent words, which have been shown to be distinctive in earlier literary studies and in Neumann's work. His results confirm some expectations, but in other respects run counter to what theological research assumes. In the case of the Pauline epistles, the first section of Romans is grouped with Colossians and Ephesians, but not all results are evident in the studies.

One of the most recent studies (Savoy 2019) tests three hypotheses: 1) that only the known four letters are from Paul; 2) that the usual seven letters are from Paul; and 3) that ten of the letters are from Paul, while 1. Timothy, 2. Timothy, Titus, and Heb are not Pauline. Based on all words that occur more than twice, three distance measures are applied to the 21 texts (Pauline letters + other letters of the Bible) and the result is shown hierarchically grouped in a dendogram. Even when changing the distance measures, there is confirmation of the four-letter hypothesis in the Pauline letters – Colossians and Ephesians were written by one author. Savoy checks his results using the so-called Imposter method (Koppel & Winter 2014), which is now considered the highest standard of stylometric technique, as it raises the distance measures in numerous iterations based on new subsets of the features, as described above. On the one hand, it is confirmed that Colossians and Ephesians were written by the same author and that the usual four form a group, but this group has a clear connection to 1. and 2. Thessalonians as well as to Philippians. On the one hand, this result goes beyond the thesis that only four letters were written by Paul, but it does not confirm the usual thesis that seven letters were written by Paul. Savoy also concludes negatively that none of the three hypotheses can be confirmed.

Those critical of the stylometric studies of Paul's letters point out that the different quantity and type of quotations from other texts could falsify the values in the case of lexis (however, some stylometricians remove the quotations from the texts), that the subsequent insertion of punctuation marks could make the use of sentence length as a feature problematic, and also that the variance could be explained by the use of different scribes. Overall, it has become clear that the stylometric studies of Paul's letters are an accurate reflection of the development of the method: from simple frequency values to univariate methods and multivariate statistics, even the criticism of the method is indicative in many ways.

Perhaps apart from the question of Q, a presumed source for the Gospels of Matthew and Luke (Poirier 2008, Oakes 2014, 153 ff., Mealand 2011), few other problems in Christian theology have attracted as much attention from stylometrics as the Pauline epistles. Nevertheless, many questions concerning the Old and New Testaments have been addressed using stylometric methods, such as whether the first and second letters of Peter really have different authors (House 2002), or whether the doubted unity of the book of Isaiah can be illuminated more precisely using stylometrics (Radday 1973).

The stylometric analysis of religious texts has also dealt with a series of questions beyond the texts of the Bible, which can only be mentioned here. For example, possible authors of the Book of Mormon have been identified (Holmes 1992, Jockers et al. 2008), there have been studies examining whether the author of the Koran is identical to the author of the Hadith, those investigating the reception tradition of the sayings and actions of the Prophet Mohammed (Sayoud 2012), studies looking at whether the Nahj Al-Balagha, which is attributed to Imam Ali, can be attributed to two authors (Sarwar & Mohamed 2022), and some studies have analyzed early Buddhist literature that came from India and was translated into Chinese (Hung et al. 2010).

Literature Cited

Bujard, W. (1973). *Stilanalytische Untersuchungen zum Kolosserbrief.* Göttingen: Vandenhoek & Ruprecht.

Burrows, J.F. (1987). *Computation into Criticism. A Study of Jane Austen's Novels and an Experiment in Method.* Oxford: Clarendon Press.

Id. (2002). 'Delta'. A Measure of Stylistic Difference and a Guide to Likely Authorship, *Literary and Linguistic Computing,* 17(3), 267–87.

De Morgan, S.E. (1882). *Memoir of Augustus De Morgan.* London: Longmans, Green. URL: http://archive.org/details/memoirofaugustusoodemouoft (Accessed: 16 June 2024).

Evert, S., Proisl, Th., Jannidis, F., Reger, I., Pielström, S., Schöch, Ch., & Vitt, Th. (2017). Understanding and explaining Delta measures for authorship attribution, *Digital Scholarship in the Humanities,* 32(2), 4–16.

Grant, T. (2022). *The Idea of Progress in Forensic Authorship Analysis.* Cambridge: Cambridge University Press.

Greenwood, H.H. (1993). St Paul Revisited. Word Clusters in Multidimensional Space, *Literary and Linguistic Computing,* 8(4), 211–219. DOI: https://doi.org/10.1093/llc/8.4.211 (Accessed: 16 June 2024).

Hirst, G., & Wei Feng, V. (2012). Changes in Style in Authors with Alzheimer's Disease, *English Studies,* 93(3), 357–370. DOI: https://doi.org/10.1080/0013838X.2012.668789 (Accessed: 16 June 2024).
Holmes, D. (1992). A Stylometric Analysis of Mormon Scripture and Related Texts, *Journal of the Royal Statistical Society. Series A (Statistics in Society),* 155(1), 91–120.
Id. (1998). The Evolution of Stylometry in Humanities Scholarship, *Literary and Linguistic Computing*, 13 (3), 111–117.
Hoover, D.L. (2021). *Modes of Composition and the Durability of Style in Literature.* New York/London: Routledge.
House, M.A. (2002). *Stylometry and the Authorship of Second Peter.* Pasadena, California: Proquest.
Hung, J.-J., Bingenheimer, M., & Wiles, S. (2010). Quantitative evidence for a hypothesis regarding the attribution of early Buddhist translations, *Literary and Linguistic Computing,* 25(1), 119–134.
Jannidis, F. (2014). Der Autor ganz nah. Autorstil in Stilistik und Stilometrie. In M. Schaffrick & M. Willand (Eds.), *Theorien und Praktiken der Autorschaft* (pp. 169–195). Berlin: de Gruyter.
Id., & Lauer, G. (2014). Burrows Delta and its Use in German Literary History. In M. Erlin & L. Tatlock (Eds.), *Distant Readings. Topologies of German Culture in the Long Nineteenth Century* (pp. 29–54). Rochester: Camden House.
Jockers, M.L., Witten, D.M., & Criddle, C.S. (2008). Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification*, *Literary and Linguistic Computing,* 23(4), 465–491.
Juola, P. (2008). Authorship Attribution, *Foundations and Trends in Information Retrieval,* 1(3), 233–334.
Id. (2015). The Rowling Case. A Proposed Standard Analytic Protocol for Authorship Questions, *Digital Scholarship in the Humanities,* 30(1), 100–113.
Kenny, A. (1982). *Computation of Style.* Oxford/Frankfurt: Pergamon Press.
Id. (1986). *A Stylometric Study of the New Testament.* Oxford: Clarendon Press.
Koppel, M., & Winter, Y. (2014). Determining If Two Documents Are Written by the Same Author, *Journal of the Association for Information Science and Technology,* 65(1), 178–187.
Ledger, G. (1995). An Exploration of Differences in the Pauline Epistles using Multivariate Statistical Analysis, *Literary and Linguistic Computing,* 10(2), 85–97.
Leech, G., & Short, M. (2007). *Style in Fiction. A Linguistic Introduction to English Fictional Prose.* 2nd ed. New York: Routledge.
Mealand, D.L. (1988). *Review of A Stylometric Study of the New Testament,* von A. Kenny. *The Journal of Theological Studies,* 39(1), 194–196.
Id. (1996). The Extent of the Pauline Corpus. A Multivariate Approach, *Journal for the Study of the New Testament,* 18(59), 61–92.

Id. (2011). Is There Stylometric Evidence for Q?, *New Testament Studies,* 57(4), 483–507.

Morgenthaler, R. (1958). *Statistik des neutestamentlichen Wortschatzes.* Zürich/Frankfurt a. M.: Gotthelf-Verlag.

Morton, A. Q., & McLeman, J. (1966). *Paul, the Man and the Myth. A Study in the Authorship of Greek Prose.* London: Hodder and Stoughton.

Neal, T., Sundarajan, K., Fatima, A., Yan, Y., Xiang, Y, & Woddard, D. (2017). Surveying Stylometry Techniques and Applications, ACM Computing Surveys, 50(6.86), 1–36. DOI: https://doi.org/10.1145/3132039 (Accessed: 16 June 2024).

Neumann, K. J. (1990). *The Authenticity of the Pauline Epistles in the Light of Stylostatistical Analysis.* Atlanta, Ga.: Scholars Press.

Nini, A. (2023). *A Theory of Linguistic Individuality for Authorship Analysis.* Cambridge: Cambridge University Press.

Oakes, M. P. (2014). Stylometric Analysis of Religious Texts. In id. (Ed.), *Literary Detective Work on the Computer* (pp. 149–206). Amsterdam: John Benjamins.

Poirier, J. C. (2008). Statistical Studies of the Verbal Agreements and Their Impact on the Synoptic Problem, *Currents in Biblical Research,* 7(1), 68–123.

Radday, Y. Th. (1973). *The Unity of Isaiah in the Light of Statistical Linguistics.* Hildesheim: H. A. Gerstenberg.

Rybicki, J. (2021). A Third Glance at a Stylometric Map of Native and Translated Literature in Polish. In M. Heydel & Z. Ziemann (Eds.), *Retracing the History of Literary Translation in Poland* (pp. 247–261). New York/London: Routledge.

Sarwar, R., & Mohamed, E. (2022). Author verification of Nahj Al-Balagha, *Digital Scholarship in the Humanities,* 37(4), 1210–1222.

Savoy, J. (2019). Authorship of Pauline Epistles Revisited, *Journal of the Association for Information Science and Technology,* 70(10), 1089–1097.

Id. (2020), *Machine Learning Methods for Stylometry. Authorship Attribution and Author Profiling.* Cham: Springer. DOI: https://doi.org/10.1007/978-3-030-53360-1 (Accessed: 16 June 2024).

Sayoud, H. (2012). Author discrimination between the Holy Quran and Prophet's statements, *Literary and Linguistic Computing,* 27(4), 427–444.

Seminck, O., Gambette, Ph., Legallois, D., & Poibeau, Th. (2022). The Evolution of the Idiolect over the Lifetime. A Quantitative and Qualitative Study of French 19th Century Literature, *Journal of Cultural Analytics,* 7(3), no pag. DOI: https://doi.org/10.22148/001c.37588 (Accessed: 16 June 2024).

Sperling, D. H. M., Kestemont, M., & Neyt, V. (2024). The Authorship of Stephen King's Books Written Under the Pseudonym "Richard Bachman". A Stylometric Analysis, *Journal of Computational Literary Studies,* 2(1), no pag. DOI: https://doi.org/10.5281/zenodo.8397200 (Accessed: 16 June 2024).

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods, *Journal of the American Society for Information Science and Technology,* 60(3), 538–556.

Literature Related

Abakuks, A. (2015). *The Synoptic Problem and Statistics.* London/New York: CRC Press.
Adams, L.L., & Rencher, A.C. (1974). A Computer Analysis of the Isaiah Authorship Problem, *BYU Studies Quarterly,* 15(1), 95–102.
Adams, L.L. (1984). A Scientific Analysis of Isaiah Authorship. In M.S. Nyman & Ch.D. Tate Jr. (Eds.), *Isaiah and the Prophets. Inspired Voices from the Old Testament* (pp. 151–164). Provo, UT: Religious Studies Center, Brigham Young University. URL: https://rsc.byu.edu/isaiah-prophets/scientific-analysis-isaiah-authorship (Accessed: 16 June 2024).
Alqurneh, A., Mustapha, A., Murad, M.A.A., & Sharef, N.M. (2016). Stylometric model for detecting oath expressions. A case study for Quranic texts, *Digital Scholarship in the Humanities,* 31(1), 1–20. DOI: https://doi.org/10.1093/llc/fqu038 (Accessed: 16 June 2024).
Barr, G.K. (2003). Two Styles in the New Testament Epistles, *Literary and Linguistic Computing,* 18(3), 235–248. DOI: https://doi.org/10.1093/llc/18.3.235 (Accessed: 16 June 2024).
Denaux, A. (2006). Style and Stylistics, with special reference to Luke, *Filologia neotestamentaria,* 19(37–38), 31–51.
Eder, M. (2013). Computational stylistics and Biblical translation. How reliable can a dendrogram be. In T. Piotrowski & L. Grabowski (Eds.), *The translator and the computer* (pp. 155–170). Wroclaw: WSF.
Elewa, A. (2019). Authorship verification of disputed Hadiths in Sahih al-Bukhari and Muslim, *Digital Scholarship in the Humanities,* 34(2), 261–276. DOI: https://doi.org/10.1093/llc/fqy036 (Accessed: 16 June 2024).
Göransson, E., Maurits, L., Dahlmann, B., Åkerman Sarkisian, K., Rubenson, S., & Dunn, M. (2023). Improved distance measures for 'fixed-content miscellanies'. An adaptation for the collections of sayings of the desert fathers and mothers, *Digital Scholarship in the Humanities,* 38(1), 127–150. DOI: https://doi.org/10.1093/llc/fqac025 (Accessed: 16 June 2024).
Grant, T., & MacLeod, N. (2018). Resources and constraints in linguistic identity performance. A theory of authorship, *Language and Law,* 5(1), 80–96.
Greenwood, H.H. (1995). Common Word Frequencies and Authorship in Luke's Gospel and Acts, *Literary and Linguistic Computing,* 10(3), 183–187. DOI: https://doi.org/10.1093/llc/10.3.183 (Accessed: 16 June 2024).
Herrmann, J.B., van Dalen-Oskam, K., & Schöch, Ch. (2015). Revisiting Style, a Key Concept in Literary Studies, *Journal of Literary Theory,* 9(1), 25–52. DOI: https://doi.org/10.1515/jlt-2015-0003 (Accessed: 16 June 2024).
Kilpatrick, G.D. (1988). Review of A Stylometric Study of the New Testament, *Novum Testamentum,* 30(4), 373–375. DOI: https://doi.org/10.2307/1560626 (Accessed: 16 June 2024).

Kinzig, Wolfram. (1990). *In Search of Asterius. Studies on the Authorship of the Homilies on the Psalms.* Göttingen: Vandenhoeck u. Ruprecht [= *Forschungen zur Kirchen- und Dogmengeschichte,* 47].

Klaussner, C., & Vogel, C. (2015). Stylochronometry. Timeline Prediction in Stylometric Analysis. In M. Bramer & M. Petridis (Eds.), *Research and Development in Intelligent Systems XXXII* (pp. 91–106). Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-25032-8_6 (Accessed: 16 June 2024).

Kocher, M., & Savoy, J. (2019). Evaluation of text representation schemes and distance measures for authorship linking, *Digital Scholarship in the Humanities,* 34(1), 189–207. DOI: https://doi.org/10.1093/llc/fqy013 (Accessed: 16 June 2024).

Koppel, M., Schler, J., & Argamon, Sh. (2011). Authorship Attribution in the Wild, *Language Resources and Evaluation,* 45(1), 83–94. DOI: https://doi.org/10.1007/s10579-009-9111-2 (Accessed: 16 June 2024).

Eid., & Winter, Y. (2012). The "Fundamental Problem" of Authorship Attribution, *English Studies,* 93, 284–291. DOI: https://doi.org/10.1080/0013838X.2012.668794 (Accessed: 16 June 2024).

Larsen, W.A., Rencher, A.C., & Layton, T. (1980). Who Wrote the Book of Mormon? An Analysis of Wordprints, *Brigham Young University Studies,* 20(3), 225–251. URL: https://byustudies.byu.edu/article/4316 (Accessed: 16 June 2024).

Libby, J. (2015). *Disentangling Authorship and Genre in the Greek New Testament. History, Method and Praxis* [Diss.]. Hamilton: MacSphere. URL: http://hdl.handle.net/11375/24538 (Accessed: 16 June 2024).

Lichtblau, D., & Stoean, C. (2023). Chaos game representation for authorship attribution, *Artificial Intelligence,* 317(103858), no pag. DOI: https://doi.org/10.1016/j.artint.2023.103858 (Accessed: 16 June 2024).

Linmans, A.J.M. (1998). Correspondence Analysis of the Synoptic Gospels, *Literary and Linguistic Computing,* 13(1), 1–13. DOI: https://doi.org/10.1093/llc/13.1.1 (Accessed: 16 June 2024).

Mealand, D. (1989). Positional Stylometry Reassessed. Testing a Seven Epistle Theory of Pauline Authorship, *New Testament Studies,* 35(2), 266–286. DOI: https://doi.org/10.1017/S0028688500024656 (Accessed: 16 June 2024).

Id. (1995). Correspondence Analysis of Luke, *Literary and Linguistic Computing,* 10(3), 171–182. DOI: https://doi.org/10.1093/llc/10.3.171 (Accessed: 16 June 2024).

Id. (1997). Measuring genre differences in Mark with correspondence analysis, *Literary and Linguistic Computing,* 12(4), 227–245. DOI: https://doi.org/10.1093/llc/12.4.227 (Accessed: 16 June 2024).

Id. (1999). Style, genre, and authorship in acts, the Septuagint, and Hellenistic historians, *Literary and Linguistic Computing,* 14(4), 479–506. DOI: https://doi.org/10.1093/llc/14.4.479 (Accessed: 16 June 2024).

Nieboer, M.C. (1970). The Statistical Analysis of A.Q. Morton and the Authenticity of the Pauline Epistles, *Calvin Theological Journal,* 5, 64–80.

Parsons, M. C., & Gorman H. M. (2012). The Assumed Authorial Unity of Luke and Acts. A review essay, *Neotestamentica,* 46(1), 139–152. DOI: https://doi.org/10.10520/EJC121513 (Accessed: 16 June 2024).

Pruscha, H. (1998). Statistical Models for Vocabulary and Text Length with an Application to the NT Corpus, *Literary and Linguistic Computing,* 13(4), 195–198. DOI: https://doi.org/10.1093/llc/13.4.195 (Accessed: 16 June 2024).

Roper, M., Fields, P. J., & Schaalje, G. B. (2012). Stylometric Analyses of the Book of Mormon. A Short History, *Journal of Book of Mormon Studies,* 21(1), 29–45. URL: https://scholarsarchive.byu.edu/jbms/vol21/iss1/4 (Accessed: 16 June 2024).

Roy, A., & Robertson, P. (2022). Applying Cosine Similarity to Paul's Letters. Mathematically Modeling Formal and Stylistic Similarities. In G. V. Allen, S. Schulthess, P. Dilley, & P. Phillips (Eds.), *New Approaches to Textual and Image Analysis in Early Jewish and Christian Studies* (pp. 88–117). Leiden/Boston: Brill [= *Digital Biblical Studies,* 5]. DOI: https://doi.org/10.1163/9789004515116_007 (Accessed: 16 June 2024).

Royal, K. (2012). Using Objective Stylometric Techniques to Evaluate New Testament Authorship, *Journal of MultiDisciplinary Evaluation,* 8(19), 1–7. DOI: https://doi.org/10.56645/jmde.v8i19.352 (Accessed: 16 June 2024).

Satos, P., & Chrysostomos, S. (2022). Examining n-grams and Multinomial Naïve Bayes Classifier for Identifying the Author of the Text "Epistle to the Hebrews". In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence,* SCITEPRESS. Science and Technology Publications, 447–458. DOI: https://doi.org/10.5220/0010970300003116 (Accessed: 16 June 2024).

Smith, M. W. A. (1988). Pseudoscience. A Comedy of Statistical Errors, *Style,* 22(4), 650–653.

Zadok, A., Zhitomirsky-Geffet, M., Schler, J., & Katzoff, B. (2023). Comparative network analysis as a new approach to the editorship profiling task. A case study of the Mishnah and Tosefta from Rabbinic literature, *Digital Scholarship in the Humanities,* 38(4), 1720–1739. DOI: https://doi.org/10.1093/llc/fqad038 (Accessed: 16 June 2024).

Network Analysis

Caitlin Burge

Abstract Over the last few decades, network analysis has grown in popularity in the humanities and presents extensive opportunities for studies in theology, offering a way to examine relational objects or entities, considering not only the importance or value of the things that are connected to one another, but the relationships themselves, and the structures they create. This chapter examines the adoption and uses of network analysis methodologies, introducing some basic terms and concepts, more broadly considering how they might be approached. It considers the appropriate questions to ask not only of the methods, but also of the data and research questions themselves before network analysis is used. In doing so, it acknowledges the criticisms and realistic precarities of the supposed concrete concepts in computational methodologies. While highlighting the complexities of network analysis, it also points to some best practice examples of different types of network analysis in theology and the wider humanities and how they might be emulated, demonstrating how network analysis may contribute fresh insights into traditional scholarly narratives.

Keywords Visualisation, Abstraction, Heuristics, Quantitative and Qualitative Methods, Historical Network Analysis

In his canonical 2002 book, *Linked,* Albert-László Barabási explored the history and prevalence of networks in the world, and how studying them opens new avenues of understanding, predicting that "network thinking is poised to invade all domains of human activity and most fields of human inquiry" (Barabási 2002, 222). His extensive work moved beyond hypothetical or random networks to develop theories and methodologies based on and suitable for real-world scenarios (ibid., 23; id. & Albert 1999). In doing so, Barabási built on work by Duncan J. Watts and Steven Strogatz, who sought to demonstrate that networks were not merely a means to explore and analyse social interactions, but all manner of connections including neural networks, power-grids, and transport systems (Watts & Strogatz 1998). Together, these publications were on the mathematical forefront of what Ruth Ahnert, Sebastian E. Ahnert, Catherine Nicole Coleman, and Scott B. Weingart term the "network turn" for humanities, a new focus and emphasis on the possibilities and value of network analysis in examining the human experience across disciplines (Ahnert et al. 2020, 3). In its simplest terms, network analysis offers a means of examining relational objects or entities, considering not only the importance or value of the entities that are connected to one another, but the relationships themselves, and the structures they create. Given the number of possibilities that this mode of analysis offers, it is unsurprising that network analysis has witnessed a surge in use not only in the natural and so-

Caitlin Burge: Network Analysis, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 203–214. DOI: https://doi.org/10.11588/heibooks.1521.c21949

cial sciences, but in the arts and humanities too, offering a new framework through which to explore traditionally humanist topics. The term *network* is not unfamiliar to scholars in the humanities and has long been used to metaphorise systems of communications and relationships; computational network analysis, then, can be used to quantify and formalise qualitative concepts that have been regularly theorised and discussed (Ahnert et al. 2020, 7; 13–24[1]).

Computational network analysis shares a variety of benefits with wider quantitative methodologies, enabling a type of *macroanalysis* or *distant reading* to view a great number of sources on a much larger scale, in ways that have not heretofore been possible (or practical) with manual or analogue means. For network analysis, this bird's eye view approach to the sources allows researchers not only to reconceptualise individual people or concepts, but entire structures of interactions, systems of relation, and societal roles. At its most fundamental, computational network analysis "makes it possible, with relative ease and speed, to measure the relationships between many entities in multiple ways, allowing a rich, multidimensional reading of complex systems never possible before" (Ahnert et al. 2020, 7).

This chapter examines the adoption and uses of network analysis methodologies, introducing some basic terms and concepts and broadly considering how they might be approached. It considers the appropriate questions to ask not only of the methods, but also of the data and research questions themselves before network analysis is used. In doing so, it acknowledges the criticisms and realistic precarities of the supposed concrete concepts in computational methodologies. While highlighting the complexities of network analysis, it also points to some best practice examples of different types of network analysis in theology and the wider humanities and how they might be emulated, demonstrating how network analysis may contribute fresh insights into traditional scholarly narratives.

1. Network Analysis – Tools, Terms, and Traps

The "ease and speed" of network analysis often relies on the relative simplicity of certain elements that are consistent across use-cases. While networks themselves can be constructed from any conceivable selection of relational entities, there are common elements of networks that can allow for comparisons across individual studies and broader fields of interests. Networks are made up of actors, objects, or concepts – known as *nodes* – and the relations between them – referred to as *edges.* While simple networks may only contain one type of node and edge, networks can be made more complex with the addition of more connection types, and one network may easily contain many different relationships at once, as we will see below. These

1 For more on networks as *thought experiments,* see ibid., 43; O'Neill 2015

types of relations may vary in their type in other ways as well, leading to *directed* and *undirected,* or *weighted* and *unweighted* networks. In an undirected network, an edge only represents whether two entities are connected, rather than accounting for characteristics such as reciprocity, order, or hierarchy; in contrast, a directed network considers the direction of a relationship. In a weighted network, values or weights are assigned to edges, e.g., strength, time, or distance. In unweighted networks, all edges are considered equal. For example, in an epistolary network of letter correspondents, each node would represent a letter writer or receiver, and the edges between them represent a set of correspondence. The network could be directed and weighted, where the *to* and *from* categories inherent in letter-writing practice are used to infer a direction in the relationship, though reciprocal correspondence would create a directed edge going in both directions. In turn, each letter contributes to the weight of an edge.

Visualisations are often the most utilised feature of network analysis. While they offer a quick overview of the dataset and relations, they can sometimes obscure as much as they reveal (cf. the chapter from J. Peters in this volume, p. 317). Katherine Bode in fact argues that graphics can hinder analysis, and "a focus on visualization impedes scholars' understanding of the evidence available to construct and interpret network models and creates perhaps insurmountable barriers to recognizing and accommodating the evidence that is absent" (Bode 2018, 125). This "hinderance" often stems from viewing these visualised networks as facts, wherein network graphs are studied as exact replicas of the source material instead of malleable representations of interactions. The arrangement of nodes and edges in a visualised network graph are an explicit design choice that more often revolves around aesthetics than any significant computational meaning; and even when they are quantitatively arranged, these often use algorithms that are not rigorously inspected or studied. In many cases, network graphs are more decorative than definitive: just as a metaphor *represents* rather than *replicates* that which it describes, so too is there an ontological gap between the visual representation of a historical period in a network and history itself (O'Neill 2015, 4–6; Ciula et al. (2018), 48; Lattmann 2018, 128 f.; 139 f.; Brughmans et al. 2016, 8). As Ahnert et al. (2020, 70) have argued, visualisation is instead "an additional means of producing, exploring, and analysing information that has proven value in both the liberal arts and the sciences." Visuals can be conducive, but we must consistently acknowledge what they can and cannot convey, using the graphs not as evidence, but as one of a number of means to explore and explain phenomena in networks.

Visualising networks offers a general overview of a network and a means to draw quick conclusions, but as Moretti (2011, 12) has argued, to gain more fruitful insight researchers should "turn away from images for a while, and let intuition give way to concepts […] and to statistical analysis." Using quantitative network measurements allows for a more in-depth understanding of networks as whole structures, as well as detailed observations about interactions and individual elements, necessitating a shift to the numerical and the greater use of measurable and quantifiable modes of analysis. Many of these modes, collectively termed *centrality measurements,* act as proxies for influence or importance, considering as the name suggests how central a node may be to the functioning of the network, and returning a number scoring their structural role based on different algorithms. *Degree* measures the number of connections (weighted or unweighted) for each node; *betweenness* establishes the likelihood that a node may act as a broker or intermediary; *eigenvector* measures the extent to which a node may influence others who are themselves highly influential; and *closeness,* quite simply, measures how close a node is to every other entity in the network. Each measurement offers a unique insight in the network and its components, and when built together can create fresh perspectives on structures of communication and connection that are more detailed than visualisations alone.

Just like network visualisations, the most effective employment of and analysis using these measures requires some understanding of the algorithms and parameters used in the calculations, which may not always be obvious in off-the-shelf software. Many available tools for network analysis cater to a range of skills, from user-interface based software or sites, such as *Gephi*[2] or *Palladio*[3], to coding packages, such as *NetworkX*[4] in *Python.* In deciding the most suitable option for a researcher or study, there is a trade-off between ease of use and computational control, and a combination of these tools may offer the best arrangement. One must bear in mind that each tool may have different defaults or limitations in their parameters, meaning that results for the same measure may differ between software, complicating analysis and replicability. When utilising these computational measures in humanistic study it is equally important to establish how they align with more traditional concepts and ideas; that is to say, how can quantitative and computational terms be translated into something more appropriate for specific humanistic research projects? In examining the potential of network analysis, the goal should not merely be to apply methodologies and theories from the social sciences to humanistic inquiries wholesale, but to create new modes of analysis that can transcend one specific topic and reconcile the two disciplines: not just adopting but *adapting* quantitative methods to suit humanistic investigations.

Though these computational measures are – tentatively – more informative than visuals alone, they still do not represent fact and are as equally open to interpretation and manipulation: quantitative results from these measures can and *should* be interrogated, challenged, and unpacked with the same scrutiny with which scholars have approached traditional humanistic sources for hundreds of years. In doing so, it is important to recognise that these measures are not only reliant on computational black-box algorithms, but the researcher's own active decision-making process

2 See https://gephi.org (Accessed: 23 June 2024).
3 See https://hdlab.stanford.edu/palladio (Accessed: 23 June 2024).
4 See https://networkx.org (Accessed: 23 June 2024).

as well. Like other forms of computational and quantitative methodology, much of network analysis is defined by sources or data collection, and how they prompt questions about core network elements. How has the network been defined and what are its boundaries? What/who is or is not included? What is considered a relationship? How have all these elements been transformed from historical or theoretical constructs into data or metadata for computational analysis? The answers to these questions, and the mode of data collection itself, may be led by the research question or the nature of the sources themselves. But some of these decisions may already be made ahead of time by how these factors – amongst others – create natural limitations; i.e., if a research project focuses on one archive, the collection itself will limit the reach of the network. For the most part, however, these are decisions that the researcher themselves must make in the process of collecting information for the network before it is even created.

By considering these questions and their impact on network analysis, Ahnert et al. (2020, 13; 75) argue that the *process* of network analysis does not begin with the network itself but rather the method of abstraction by which sources are transformed into quantitative networks. This process, whereby researchers unpack the different layers of information in qualitative sources, “requires a prior *mental* manoeuvre of translating cultural artefacts into an abstracted form to see whether they are compatible with the input requirements of the available tools” (ibid., 75). By actively acknowledging and engaging with this process of abstraction, we can more critically consider how sources are adapted for network analysis and how network analysis fits a particular set of sources, considering what is and what is not included. Importantly, this must be understood as a continuum: even at the strictest level of abstraction, the dropped information is not ignored indefinitely, but is merely not in use right now, much like qualitative selection and analysis of sources (ibid., 51).

This process is not just a question of what is abstracted, but *how* the sources are transformed. Criticism has focused on this supposed inability in network analysis to grasp the more complex ideas of humanities disciplines; as Elwert (2020, 182) critiqued in his survey of network analysis in religious studies, this approach “tends to reduce religious phenomena to social processes but neglects the content of religious exchange… [which] might in part explain why network analysis has been adopted rather reluctantly in main-stream religious studies.” This assessment is problematic, however, placing responsibility for these *failures* on the methodology itself, rather than how it has been applied. To use computational network analysis to its full advantage, greater effort and active participation in the abstraction process is required, reflecting on how the transformation from concepts and constructs to quantifiable connections is, ultimately, defined by the researcher. This abstraction offers an important and iterative means to more concretely examine and conceptualise sources and data and their meaning or value in scholarly narratives. Paying more critical attention to this process and what it means for networks counters any blanket rejections of network analysis and suggestions that it does not suit theology and religious studies.

The formation of networks and use of computational analysis therein relies not only on abstraction but, as with any other scholarly practice, the original selection of sources and data as well. Networks, then, are much like archives: they are not organic, naturally occurring entities but carefully curated collections subject to layers of selection, from the original writers or creators through a multitude of readers, archivists, and scholars up to today. Acknowledging these layers and influences not only improves understanding of the analysis and its results, but often the sources themselves too. While this type of critique and recognition in methodology applies across computational methods in the humanities, it is especially important when using network analysis to examine relational objects to acknowledge our own, somewhat hierarchical, relationship with the network as well.

2. Best Practice and Best Examples

The fact remains that, when used effectively, combining the qualitative and quantitative allows for a more nuanced and well-rounded understanding of topics, in which elements of both disciplines can be used simultaneously: close *and* distant reading, interpretative *and* descriptive work. As the remainder of this chapter turns to examples of different types of network analysis in the humanities, it is important to consider how this nuanced application forms ideas of best practice: contributing both to understandings of computational practice and traditional scholarly narratives, while reflecting critically on the process and discerning application of appropriate methodologies.

One of the most popular forms of network analysis in the humanities is *social network analysis* or SNA, examining networks primarily built on either pre-existing evidence of interactions or recorded from interviews and observation, and is the approach most commonly found under the umbrella term *historical network analysis.* In studying letter metadata – information *about* documents, rather than the contents of sources themselves – Ahnert & Ahnert (2015) utilise this approach to examine underground networks of Protestant communities in the reign of Catholic Mary I of England. Quantitative network analysis of epistolary networks in this period unsurprisingly confirmed some expected actors of importance, in particular prolific martyrs. But having identified these actors, the measures were also used to construct network fingerprints to find potentially unknown or unexpected actors in similar roles, creating replicable experiments not only for this singular dataset but beyond as well. In using computational measurements to identify different types of interaction profiles in the network, Ahnert and Ahnert demonstrate how quantitative network analysis employed in tandem with extensive traditional historical knowledge to understand how these network roles translate into early modern realities can offer nuanced insight into relational structures.

While similarly investigating historical networks, Düring's (2016) exploration of Jewish support networks during the Second World War uses a variety of relationship categories, including "form of help, intensity of relationships, motives for action, date of help and date of first meeting" based on primary first-hand accounts. In doing so, Düring examines not only the *existence* of edges but their explicit role in the societal structure and how this may indicate certain nodes or relationships of importance. By investigating the history of the Segal family as a case study, network analysis is used to examine brokerage relations providing aid. As such, Düring applies computational methodologies as a formal mode to both qualify and quantify hypothesised roles in historical interactions.

While Ahnert and Ahnert utilise metadata to build social networks, Düring extracts the interactions from the available texts themselves, an approach Bourke (2024) similarly adopts in his examination of women writers in John Locke's correspondence. Acknowledging that the correspondence of John Locke fails to offer metadata "comprised of diverse senders and recipients," Bourke instead makes use of the contents of the letters as well, including instances of citation – where a writer mentions another person – and co-citation – where two people who are not the writer are referenced alongside one another – to build additional networks of social interaction. This is to say, where Düring extracts network edges out of text that explicitly *describes* social interactions, Bourke's citation and co-citation networks replicate *inferred* connections, layering metadata and network constructs with careful close reading. In doing so, Bourke (2024) "map[s] the intellectual and social structure of the conversations Locke was having within his correspondence," critically offering space to (namely female) actors that otherwise hold little sway in a metadata-only network, and examining how this may indicate other actors of influence.

One of the more popular uses of these co-citation networks in religious studies is examining relational structures surrounding Jesus in the Bible, again extracting data for network edges from relationships both described and inferred in the text. McClure's (2020) paper goes beyond utilising these structures as a means to explore Jesus' social network as one singular entity, instead constructing different networks from the four Gospels of the New Testament and comparing results to consider how network analysis may contribute to more traditional examinations of literary and textual overlaps and differences between the Gospels. This study both confirms already identified points of overlap or differentiation, but also identifies new points of interest – in particular around the inclusion of women and stigmatized people – that may have been acknowledged in a textual context, but not a relational structure one. McClure's investigation is primarily based on pre-existing narratives, and she acknowledges that the findings are likely unsurprising to biblical scholars, but argues that the study "supplements their textual, historical, and theological observations by exploring relational and structural patterns not previously examined" (ibid. 47), effectively demonstrating how network analysis can still offer interesting insights into scholarly debates with a long and rich history.

These studies primarily adopt close-reading techniques to extract information for social network analysis, but networks can also be employed as means of text analysis in and of themselves. While social network analysis primarily focuses on social relations between objects or, as in these case studies, people, similar structures can be used to study singular or canons of text, producing similar insights to other forms of digital text analysis such as topic modeling (cf. the chapter from M. Althage in this volume). This may involve word association or semantic networks, which attempt to replicate mental representations and understandings of linguistic connections, but it may also resemble something similar to the co-citation networks, using co-occurrences of words in a sentence or paragraph to construct networks (Czachesz 2016, 43[5]). Like Jennifer M. McClure, István Czachesz also employs network analysis to examine differences and similarities between Gospel stories, but instead using these word co-occurrences to consider literary and linguistic differences, rather than social ones. In doing so, Czachesz critically examines how ideas or elements are emphasised differently across iterations of miracle stories. Though methodologies such as these offer simplistic findings when applied on a small scale, Czachesz's study acts as a model of the possibilities of these approaches when applied to a much larger textual corpus.

The use of network analysis to explore literary and linguistic connections is just one demonstration of how nodes may be made up of any relational entity. With this in mind, *Actor-Network Theory* (ANT) – proposed by Bruno Latour in the 1980s – suggests that everything in the world exists in one network together, including humans and non-humans, both animate and inanimate, extending as much to ideas as to physical objects and beings (Latour & Woolgar 1986; Latour 2005; Van Oyen 2016). This theory acknowledges that ideas and objects are not just connected to one another, but are able to influence and impact one another, as much as nodes in a social network might. Checketts (2017) utilises this theory as a new framework examine and unpack the relationship between Christianity and technology, and the evolving place of both in the modern world. While Checketts' argument remains qualitative rather than quantitative – choosing not to evoke the digital and technological mediums he discusses – he performs the important manoeuvres discussed above to translate theological debates into network theory, setting the stage for possible computational measures in a further study.

While best practice is still developing around these different types of networks and their varied applications in the humanities, we can posit how layering these approaches might provide the most fruitful insights. Take, for example, the development and spread of ideas across Europe during the Reformation. Several state papers contain evidence of interactions between important rulers, theologians, and religious leaders, and building networks out of these may allow us to consider structures of influence and power. These already offer an interesting framework to examine inter-

5 For further examples see Purschwitz 2018; Sangiacomo et al. 2022.

actions using social network analysis methodologies; but how could this be layered further with other modes of network analysis? What if these epistolary metadata networks were supplemented with citation or co-occurrence networks built with the *contents* of the letters? Or, to take it another step, moving closer to Latour's *Actor-Network Theory,* what if the nodes were made of both correspondents and the *ideas* in the letters or wider writings, as in semantic networks? Can we model connections between language choices in tracts or translations, the theologians who wrote them, and the evolution of religious change in the early modern world? While these remain only hypothetical questions, it prompts thinking as to the possibilities of multi-layered and multi-dimensional networked thinking in religious and theological studies.

Though network analysis may also only go as far to confirm actors and entities of importance or structures that are already confirmed in traditional scholarly literature, this holds value in validating supposedly new discoveries that may also appear. By employing these methods as new frameworks of understanding rather than new types of *evidence,* we can adjust expectations of what computational network analysis will achieve, and therefore how useful it is. In doing so, we can also prompt new means of approaching questions around importance or influence and redefine how we conceive of connected structures throughout theology and religious studies. The interaction between quantitative and qualitative mentalities is therefore an iterative, heuristic one. Though some elements of traditional qualitative research may resist the idea that aspects of the humanities can be quantified, it is undeniable that these methods offer modes of measurement and, more importantly, *comparison,* that have not so easily been achieved before. While it is important to recognise the malleability and even *fallibility* of data and the structures they create, by acknowledging and *incorporating* this awareness into our approach, computational network analysis can offer exciting and thought-provoking outcomes, and when used together with traditional approaches, can create more developed research processes and enriched narratives in theology, religious studies, and beyond.

References

Ahnert, R., & Ahnert, S. E. (2019). Metadata, Surveillance and the Tudor State, *History Workshop Journal,* 87, 27–51. DOI: https://doi.org/10.1093/hwj/dby033 (Accessed: 23 June 2024).

Eid. (2015). Protestant Letter Networks in the Reign of Mary I. A Quantitative Approach, *ELH,* 82, 1–33. URL: https://www.jstor.org/stable/24477807 (Accessed: 23 June 2024).

Eid. (2023). *Tudor Networks of Power.* Oxford: Oxford University Press.

Eid., Coleman, C.N., & Weingart, S.B. (2020). *The Network Turn. Changing Perspectives in the Humanities.* Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781108866804 (Accessed: 23 June 2024).

Barabási, A.-L. (2002). *Linked. The New Science of Networks.* New York: Penguin Group.

Id., & Albert, R. (1999). Emergence of Scaling in Random Networks, *Science,* 286(5439), 509–512.

Bode, K. (2018). *A World of Fiction. Digital Collections and the Future of Literary History.* Ann Arbor: University of Michigan Press.

Bourke, E. (2024]). The Reception of Women Letter-writers in the Correspondence of John Locke (1632–1704), *Huntington Library Quarterly Special Issue. Network Analysis and the Early Modern Archive,* 86(2), 227–256.

Brughmans, T., Collar, A., & Coward, F. (2016). *The Connected Past. Challenges to Network Studies in Archaeology and History.* Oxford: Oxford University Press.

Ciula, A., Eide, Ø., Marras, C., & Sahle, P. (2018). Introduction to Historical Social Research, *HSR-Supplement,* 31, 7–29.

Checketts, L. (2017). The Cross and the Computer. Actor-Network Theory and Christianity, *Theology and Science,* 15(1), 116–127.

Czachesz, I. (2016). Network Analysis of Biblical Texts, *Journal of Cognitive Historiography,* 3, 43–67.

Drucker, J. (2011). Humanities Approaches to Graphical Display, *Digital Humanities Quarterly,* 5(1), 1–52. URL: http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html (Accessed: 23 June 2024).

Düring, M. (2016). The dynamics of helping behaviour for Jewish fugitives during the Second World War. The importance of brokerage. The Segal Family's Case, *Mass Violence & Résistance.* Online. URL: https://www.sciencespo.fr/mass-violence-war-massacre-resistance/en/document/dynamics-helping-behaviour-jewish-fugitives-during-second-world-war-importance-brokerage-se.html (Accessed: 23 June 2024).

Elwert, F. (2020). Social and semantic network analysis in the study of religions. In F. Kerschbaumer, L. von Keyserlingk-Rehbein, M. Stark & M. Düring (Eds.), *The Power of Networks. Prospects of Historical Network Research* (pp. 172–186). London: Routledge.

Emirbayer, M., & Goodwin, J. (1994). Network Analysis, Culture and the Problem of Agency, *American Journal of Sociology,* 99, 1411–1454. DOI: https://doi.org/10.1086/230450 (Accessed: 23 June 2024).

Greteman, B. (2021). *Networking Print in Shakespeare's England. Influence, Agency, and Revolutionary Change.* Stanford: Stanford University Press.

Klein, L.F. (2013). The Image of Absence. Archival Silence, Data Visualization, and James Hemings, *American Literature,* 85(4), 661–688.

Latour, B. (2005). *Reassembling the social. an introduction to actor-network-theory.* Oxford: Oxford University Press.

Id., & Woolgar, S. (1986). *Laboratory life. The construction of scientific facts.* Princeton: Princeton University Press.

Lattmann, C. (2018). Iconizing the Digital Humanities. Models and Modeling from a Semiotic Perspective, *Historical Social Research. Supplement,* 31, 124–146. DOI: https://doi.org/10.12759/hsr.suppl.31.2018.124-146 (Accessed: 23 June 2024).

McClure, J. M. (2020). Jesus's Social Network and the Four Gospels. Exploring the Relational Dynamics of the Gospels Using Social Network Analysis, *Biblical Theology Bulletin,* 50(1), 35–53. DOI: https://doi.org/10.1177/0146107919892841 (Accessed: 23 June 2024).

Moretti, F. (2011). Network Theory, Plot Analysis, *LiteraryLab,* 2, 1–12. URL: https://litlab.stanford.edu/assets/pdf/LiteraryLabPamphlet2.pdf (Accessed: 23 June 2024).

Nagel, A.-K. (2012). Measuring the Relational. How to collect data on religious networks. In L. Berzano & O. Riis (Eds.), *New Methods in the Sociology of Religion* (pp. 181–205). Leiden: Brill Academic Publishing [= *Annual Review of the Sociology of Religion,* 3].

Id. (2022). Network Analysis. In S. Engler & M. Stausberg (Eds.), *The Routledge Handbook of Research Methods in the Study of Religion* (pp. 375–386). London: Routledge.

O'Neill, L. (2015). *The Opened Letter. Networking in the Early Modern British World.* Philadelphia: University of Pennsylvania Press.

Perry, B. L., Pescosolido, B. A., & Borgatti, S. P. (2018). *Egocentric Network Analysis. Foundations, Methods, and Models.* Cambridge: Cambridge University Press.

Purschwitz, A. (2018). Netzwerke des Wissens. Thematische und personelle Relationen innerhalb der halleschen Zeitungen und Zeitschriften der Aufklärungsepoche (1688–1818), *Journal of Historical Network Research,* 2, 109–142. URL: http://jhnr.uni.lu/index.php/jhnr/article/view/47 (Accessed: 23 June 2024).

Rollinger, Ch., Düring, M., Gramsch-Stehfest, R., & Stark, M. (2017). Editors' Introduction, *Journal of Historical Network Research,* 1, i–vii. URL: http://jhnr.uni.lu/index.php/jhnr/article/view/19 (Accessed: 23 June 2024).

Sangiacomo, A., Tanasescu, R., Hogenbirk, H., & Donker, S. (2022). Recreating the Network of Early Modern Natural Philosophy. A Mono- and Multilingual Text Data Vectorization Method, *Journal of Historical Network Research,* 7, 33–85. DOI: https://doi.org/10.25517/jhnr.v7i1.129 (Accessed: 23 June 2024).

Van den Heuvel, Ch. (2015). Mapping Knowledge Exchange in Early Modern Europe Intellectual and Technological Geographies and Network Representations, *International Journal of Humanities and Arts Computing,* 9(1), 95–114.

Van Oyen, A. (2016). Networks or Work-Nets? Actor-Network Theory and Multiple Social Topologies in the Production of Roman Terra Sigillata. In T. Brughmans, A. Collar & F. Coward (Eds.), *The Connected Past. Challenges to Network Studies in Archaeology and History* (pp. 35–56). Oxford: Oxford University Press.

Watts, D.J. (2002). A simple model of global cascades on random networks, *Proceedings of the National Academy of Sciences of the United States of America*, 99, 5766–5771. DOI: https://doi.org/10.1073/pnas.082090499 (Accessed: 23 June 2024).
Id., & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks, *Nature*, 393(6684), 440–442.

Discourse Analysis

Alexander Lasch

https://orcid.org/0000-0003-4046-3993

Abstract This article presents the basic assumptions of linguistic discourse analysis, particularly regarding how the constructivist understanding of language relieves interpretations of the contextual embedding of linguistic and discursive actions. The method is to focus on practices of machine analysis of discourse. The article analyzes contemporary research approaches (metaphor and frame semantics) combining both quantitative and qualitative approaches.*

Keywords Discourse Analysis, Discourse Linguistics, Metaphor, Frame Semantics

1. Reconstruction of Discourse Formations

1.1 Language, Knowledge, and Society

> The concept of discourse is one of the most powerful guiding concepts in modern humanities, cultural, and social sciences (Gardt 2017, 2).

This brief article cannot fully explore the "powerful" and dazzling aspects of the concept of discourse (cf. Göhring 2023, 95–202). Rather, the aim here is to trace a specific conception of discourse as established in discourse linguistics in German Studies (cf. Felder & Jacob 2017). Discourse analysis is characterized by a constructivist understanding of language (Felder & Gardt 2018), the operationalization of which will provide a key element to the current article (Section 2). Fritz Hermanns (1995) posits an apt starting point for this line of argumentation. He asserted that the mentality, "the totality of the collective thinking, feeling, and willing" of a community, could be reconstructed linguistically (Hermanns 1995, 89). Inherent in this assumption is the conviction that conclusions are drawn from the way in which a society linguistically negotiates *about* society and how this society *conceptualizes* specific forms of social interaction, power structures, bodies of knowledge, economic contexts, transcendence, mortality, etc. and *constructs* specific social realities by means of language.

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

Alexander Lasch: Discourse Analysis, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 215–228. DOI: https://doi.org/10.11588/heibooks.1521.c21950

Following Michel Foucault, both discourse linguistics in German Studies and critical discourse analysis (CDA) in Germany (with international precursors) have dedicated themselves to analyzing these functional relationships since the 1980s. While discourse linguistics in German Studies attempts to grasp the descriptive connection between language, knowledge, and society with regard to cultural, historical, political phenomena, and over the years has turned to cognitive linguistic research perspectives (cf. Spieß & Köpcke 2015 on *Metonymy and Metaphor*), the CDA is primarily concerned with uncovering the connection between power and ideology on the basis of linguistic analyses to show how social hierarchies are reproduced through language, which is always used ideologically to maintain established power structures, particularly in a stage of globalized capitalism (for an introduction, cf. Jäger 2015, for a critical appraisal, see Niehr 2014, 51 f.). Busse, an important representative of discourse linguistics in German Studies, rightly sees that

> this needs to be emphasized in the present context – that the category of 'power' is fundamentally not an ideological and evaluative one, but is in principle descriptive (and was also meant as such by Foucault, who introduced it into discourse analysis, see Busse 2013, 35 f.) (Kämper 2017, 265).

The brief differentiation between discourse linguistics in German Studies and CDA suffices to show that the focus of this article is the German Studies' discourse linguistic approach; nevertheless, the CDA approach will help develop (disruptive) potential in the theological, church-historical, or general religious studies context. However, it remains

> a fundamental problem that many critics have pointed out […] that many representatives of CDA do not put into practice the claim they have formulated themselves to reflect critically on their own point of view […]. These studies [run] the risk of projecting a result assumed for the time being based on preferred social models onto the data and thus only ever confirming their own views themselves (Spitzmüller & Warnke 2011, 112 f.).

All forms of discourse analysis rely on both qualitative and quantitative methods (Section 2.2). The methods are both classic hermeneutic methods of textual content and argumentation analysis, such as analyzing political speeches or debate contributions in national newspapers as well as statistical evaluations of machine-readable collections of spoken and written language. What is important here is that the researchers always relate their research question, the research material, and thus the subject matter closely to one another and increasingly adapt them in iterations, so that discourse analyses are characterized by a rich set of methodological instruments and are open to theoretical and methodological expansion, which is essentially a consequence of the version of the concept of discourse coined by Dietrich Busse in

Discourse Semantics (1987) in contrast to Foucault's discourse analysis. Together with Wolfgang Teubert, Busse asked as early as 1994 whether a discourse was a linguistic object. At that time, the potential of the term *discourse* could not yet be assessed. Currently, many linguistic discourse studies following Busse and Teubert show that the options offered by Foucault's extension of the term were highly adaptable. For example, the underlying Foucauldian idea of knowledge formation being embedded in language was adopted in the concept of discourse proposed by Busse and Teubert. At the same time, their model emphasized how language also produces knowledge formations. There is more conceptual work to be done here: The concept of discourse (Section 1.2) also establishes an understanding of the corpus (Section 1.3), which generations of scholars have worked on, and which will be discussed in more detail, as it has far-reaching consequences for possible forms of analysis. The same applies to Fritz Hermanns' supplementation and expansion of this understanding of the corpus. Unlike Busse and Teubert, Hermanns no longer focused on the *relevant change of discourse* in the texts as a relevant criterion for corpus creation. The focus had been proposed to demonstrate (telic) discourse development by analyzing the texts that shape and change a discourse. Hermanns, on the other hand, pointed (implicitly) to the *relevant confirmation of discourse* of texts and their significance for a history of mindsets as the "royal way of language history" (Hermanns 1995, 71) (Section 1.3). Methodologically, this differentiation opens two historical forms of discourse-linguistic analysis. The studies emphasizing the *relevant change of discourse* are often qualitative and make use of *traditional* hermeneutic methods of interpretation. However, if *relevant confirmation of discourse* is used as a criterion for creating the body of data, quantitative studies are to be expected, the results of which are passed on to qualitative analysis (Section 2).

1.2 Discourse

> From the very beginning, the fundamental codes of a culture that govern its language, modes of perception, interactions, technology, values, and practical hierarchy, set the empirical orders for every person, with which one will have to deal, and in which one finds oneself (Foucault 1996, 10).

In *The Order of Things,* Michel Foucault developed a concrete research program to analyze (historical) knowledge formations and their conditions of possibility by uncovering cultural sediments, to which he also counted textual traditions and contexts of statements as "fundamental codes". In the *Archaeology of Knowledge,* Foucault then refers to contexts of statements and their networking as a "discourse [which is] a set of statements that belong to the same formation system" (Foucault 1981, 156). At the end of the 1980s, Busse associated Foucault's concept of discourse to "discourse semantics" (Busse 1987, 251). Discourse semantics is interested in the reconstruction of "knowledge" and the "conditions of knowledge," which elude the conscious access of the speaker (Busse 1987, 256). Discourses are constituted by communicative actions embedded in a concrete context (Busse 1987, 259), in which, along with Foucault, the knowledge of a culture is, to a certain extent, sedimented. However, it is not possible to deduce knowledge formations and their conditions exclusively from these sediments. Busse also points to the pragmatic dimension of action in language and to the fact that language itself constitutes knowledge. Busse thus focuses on the

> scenario of the collective knowledge of a given discourse community in a given period with regard to the thematic area or field of meaning or discourse formation chosen as the object of investigation (Busse 1987, 267).

Busse and Teubert established a research program on this basis, which (with Pêcheux) is oriented entirely towards the analysis of texts:

> By discourses we mean, in a practical research sense, virtual text corpora whose composition is determined by content-related (or semantic) criteria in the broadest sense (Busse & Teubert 1994, 14).

For *practical research reasons,* Busse and Teubert do not take up Foucault's concept of the *dispositif,* which is once again gaining importance in the implementation of multimodule approaches in discourse linguistics. Instead, they concentrate exclusively on what linguistics had to investigate in the 1990s: Texts. These clarifications (knowledge constitution of language, grouping of statements and contexts in a virtual text corpus) proved to be extremely fruitful, for which Busse & Teubert could only hope (1994, 13; cf. parenthetically the contributions in Warnke 2007). But what exactly is the virtual corpus?

1.3 Corpus

A virtual corpus includes *all* texts that demonstrate semantic relationships to one another, i.e., they are related in terms of content. To be an object of research, however, they must satisfy specific qualifications (time period, territory, sector of society, text type, area of communication, etc.) (cf. Busse & Teubert 1994, 14). If *all* texts on a specific topic were examined, the scope of virtual corpora in the 1990s would have pushed all studies to the limits of feasibility, given that until the end of the 2010s, quantitative methods in discourse linguistics were still the exception rather than the rule (cf. Bubenhofer 2009). Busse & Teubert refer to the pieces of virtual corpus as the concrete corpus, which forms the basis for a discourse analysis. These parts of the virtual corpus are the "subsets of the respective discourses" (Busse & Teubert 1994, 14) not formed based on quantitative (representativeness), but qualitative criteria ("relevance criteria"), which researchers define based on their own research interests. Along with the limitations mentioned above, the availability of sources and their *relevance to changes in discourse* must be taken into account:

> Redundancies [are to be avoided] and primarily those texts [are to be included] that have significantly influenced the structure and course of the discourse [...]. [The] constitution of the discourse [...] thus always presupposes interpretative actions on the part of the researchers (Busse & Teubert 1994, 16).

The interpretive background of the researchers must be constantly and critically examined as the concrete corpus must prove to be a meaningful section of the virtual corpus: Exploration of the virtual corpus and continual specification of the concrete corpus as well as the adaptation of the questions in the constitution of the object are among the basic operations of discourse linguistics. The *discourse* is thus *de facto* the *object, method,* and *result* of the investigation. Consequentially, Busse and Teubert privileged the criterion of *relevant change of discourse* when compiling the concrete corpus at the expense of the criterion of *relevant confirmation of discourse.* In so doing, they (initially) excluded studies dealing specifically with stable structures, which provide the only framework by which phenomena of change can be adequately analyzed (cf. the works of the French school of Annales, especially Braudel 1977, 50f.). For serial, seemingly uniform bearers of tradition can be used to observe new narrative postures, reinterpretations of ritualized linguistic patterns, expansions and narrowing of conceptual meanings, thematic reorientations etc. particularly well – albeit no longer qualitatively in a concrete corpus compiled by hand. Hermanns extended Busse's (1987) discourse semantics that tended in this direction. Hermanns questioned the effects it might have on the histories of institutionalized and narrated language if historical semantics and textual linguistics were combined into the "paradigm of conceptual history" (Busse 1987 in Hermanns 1995, 80). He proposes the idea of the history of language as the history of mentality (Hermanns 1995, 70). Relatedly, he speaks of discourse semantics as "linguistic anthropology" (Hermanns 1994) and later develops the notion of a "socio-pragmatic historiography" (Hermanns 2001, 596). The basic idea is to understand the use of language in a linguistic community as an indicator of collective thinking, feeling, and will of the community, in the sense of Foucault, as a mentality, which comprises "1) the totality; 2) of habits or dispositions; 3) of thinking; 4) of feeling; and 5) of willing or wanting in 6) social groups" (Hermanns 1995, 77). Hermanns not only extends Busse and Teubert's definition of the corpus to include the imaginary corpus (Hermanns 1995, 89), which is a hypothetical quantity encompassing all thematically related texts, yet not fully accessible for practical research. He also calls for an examination of "what is serial, what is usually generally in the group and therefore in the sources" (Hermanns 1995, 89). Along with others, Hermanns argues that *discourse confirming* texts should also be included in discourse linguistic analyses, since the *discourse changing* quality of texts can only be recognized and described against the background of the serial. Busse also gradually expanded the research program, moving once again in 1997 closer to Foucault. Unlike Teubert, who continued to operationalize discourse strictly as a set of medially written texts (cf. Teubert 2006), Busse was primarily interested in knowledge formations (2007), which were a prerequisite for the genesis of texts and other cultural testimonies, and which he sought to describe as “discourse semantic basic figures” (cf. also Busse 1997 and 2000). Busse mainly understood these basic figures, which will be formative for discourse linguistics, as (textual) semantic features of statements and complexes of statements such as thematic developments, isotopies and isotopic chains, topoi, presuppositions, specific keywords, terms, etc., which are characteristic of a discourse.

2. Perspectives of Analysis

All these proposals were introduced almost simultaneously into the field of research. It is inaccurate to describe these as “extensions” of the concept of discourse. Rather, they are a series of distinct yet interrelated approaches that were further developed and evaluated in subsequent studies over time. Combined with the consolidation of what a *discursive linguistic after Foucault* (Warnke 2007) in fact is, several stages can be observed in the research that served to operationalize the understanding of discourse further. These works are above all in connection with the activities of the research network *Sprache und Wissen* (cf. Felder 2008 as well as Felder & Müller 2009), the publication series by the same name[1], and documented in the *Handbuch Sprache und Wissen* (cf. Felder & Gardt 2015). The results of the work are accessible through innovative online resources such as the *discourse monitor.*[2]

2.1 Characteristics of Established Objects

A crucial momentum for the particular characteristic of discourse linguistic studies is Hermanns’ idea of analyzing “key words,” which are not only the “vehicle or ciphers of thought,” but “indicates, while also propagating, the aim or a program *as* the aim (program) (Hermanns 1994, 12). Wengeler, whose work on *Topos und Diskurs* (2003) directly correlates with Hermanns’ key words research, emphasizes in 2017 that it was not only Hermanns’ achievement to point out the relevance of ‘key words,’ but also to introduce categories into the discussion with a systematization understood as

1 See https://www.degruyter.com/serial/suw-b (Accessed: 17 June 2024).
2 See https://diskursmonitor.de (Accessed: 17 June 2024).

indicators of *relevant discourse change* and correlated with various arguments and negotiating positions in discourse:

> In a variety of ways, Fritz Hermanns has most clearly and systematically addressed not only the cognitive dimensions of lexical, particularly political semantics. He also directed attention to the appellative, emotive, and volitive dimensions. Terminologically, the differentiation of political catchwords into *flag words, stigma words, high-value* and *low-value words,* and *affirmative words,* all go back to him (Hermanns 1994) (Wengeler 2017, 25).

Contradictory, or at least strongly varied, contours of social reality, the study of which is inherent in the program of linguistic discourse in German Studies from the outset, can be observed well when negotiated publicly – linguistic discourse studies are thus characterized using easily accessible, mostly public, written sources. Alongside the debate on the type of *discourse* that is actually being depicted (Roth 2015), namely, the discursive positions represented in mass media or on the stages of parliamentary democracy, and not *the* use of language, also became apparent that the relationship of the criteria of the *relevance of the change in discourse* and *relevance of the confirmation in discourse* had to be adjusted under the category of a *corpus pragmatic* when large machine-reading corpora became available from the beginning of the 2000s (cf. Bubenhofer 2009) and quantitative analyses began to compete with qualitative analyses (Felder et al. 2012a; b).

The two characteristics, (1) analyzing divergent drafts of social reality on the basis of (2) specific publicly accessible corpora, so shaped by research in discursive linguistics that dealing with "semantic struggles" has become one (!) of the distinguishing features of discursive linguistic research (cf. also *Agonalität* in the title of Göhring 2023). One can exemplify this relation by looking at the anthology *Agonalität* edited by Ekkehard Felder (2006a), whose essay focuses on these negotiations:

> It is considered fundamental that knowledge, in all its fields, is constituted, among other things, by language. In this respect, the "fight over the subject matter" can be a "fight over words", in short, a "semantic battle". This semantic battle can arise in three different ways: as a battle over appropriate terms, as a battle over the meaning or aspects of meaning, and/or as a battle over the facts that are only formed and shaped, i. e. fixed, regarding expressions (Felder 2006b, 1).

2.2 Methods

A discussion on the methods of discourse analysis cannot be done without the "Discourse Linguistic Multi-Layer Analysis" (DIMEAN) proposed by Spitzmüller & Warnke (2011). DIMEAN does not present an independent concept but is (1) an attempt to order objects and methods that have been used for discourse analysis in the polyphonic discourse linguistic research according to Busse and (2) an effort to make the results of the discourse linguistic studies intersubjective, reliable, and transparent, as well as to offer the opportunity for critical reflection on the limitations of the investigations. DIMEAN aims at describing complex discourses on various systematic levels of language, which are ultimately used for the interpretation of discursive practices, discursive positions of actors and discursive rules (cf. Fig. 1 on the next page).

The theoretical models of text linguistics, conversation analysis (interactional linguistics), sociolinguistics, such as historical semantics (in the sense of Busse, Teubert, and Hermanns) all become integrated on the *intratextual layer.* On this level, quantitative methods are gradually applied (above all through keyword and collocation analysis) to determine key words, identify specific syntactic structures, ascertain lexical lines of opposition, and describe lexical fields. The *actors layer* and the *transtextual layer* remain in the purview of qualitative description for a long time to come. However, even nowadays, machine methods are being employed to form hypotheses (Topic Modeling).

Based on the considerations of the specific understanding of discourse and corpus (section 1) and the explicit characteristics of discourse linguistic studies (section 2.1), it is now evident which quantitative methods are particularly relevant: collocation analysis and key word analysis (cf. Bubenhofer 2009), increasingly in recent years collostructional analysis (cf. Stefanowitsch & Gries 2009, esp. Stefanowitsch 2020, among others), sentiment analysis, and (word embedding in) topic modeling, which, among other aspects (Named Entity Recognition) are presented in detail in the following chapters of this volume. The volume *Forschen in der Linguistik* (Beißwenger et al. 2022) provides an excellent overview, not only in the collected case studies, such as Heidrun Kämper's essay (2022) "Emotion und Spachgebrauch: Ein linguistischer Beitrag zur Entstehungsgeschichte des Nationalsozialismus," which directly identifies aspects of current research interests. The volume also describes the complete methodological repertoire useful in discourse linguistics in German Studies through focused essays on "Daten – Metadaten – Annotationen," "rechtliche und ethische Aspekte beim Umgang mit Sprachdaten," "Erhebung und Aufbereitung von Spachdaten," "Korpusressourcen zum Deutschen," and esp. "Werkzeuge für die empirische Sprachanalyse" comprised of machine supported transcription of spoken language (Schmidt 2022), statistical analysis (Wolfer & Hansen 2022), corpus analysis (Haaf 2022), automatic language analysis (Horbach 2022), and possibilities of collaborative annotation (Fladrich 2022).

Transtextual Layer	Discourse-based analysis	Intertextuality, schemata, basic discursive semantic figures, topoi, social symbolism, indexical orders, historicity, ideologies/mentalities, general social and political debates	
Actors	**Discourse shaping** ↓ ↑ **Rules of discourse**	Interaction roles	Author, anticipated addressees
		Discourse positions	Social stratification/power, discourse communities, ideology brokers, voice, verticality status
		Mediumship	Medium, forms of communication, areas of communication, textual patterns
Intratextual Layer	Text-based analysis	Visual text structure	Layout/design, typography, text-image relationships, materiality/text carrier
		Macrostructure: Text theme	Lexical fields, metaphor fields, lexical opposition lines, thematic development, text strategies/text functions, text type
		Mesostructure: Themes in subtexts	
	Proposition-based analysis	Microstructure: Propositions	Syntax, rhetorical figures, metaphorical lexemes, social, expressive, deontic meaning, presuppositions, implicatures, speech acts
	Word-based analysis	Multiple-word units	Keywords, stigma words, names, ad hoc formations
		One-word units	

Fig. 1 Model of Discourse Linguistic Multi-Layer Analysis (DIMEAN)

3. Prospect

The above outline of the concepts of discourse and corpus, the criteria for the relevance of change and confirmation in discourse, as well as the focus on the analysis of divergent proposals of social reality based on specific publicly accessible corpora of discourse has clearly established that the discourse analysis in German Studies profits greatly from the progress of digitalization of its sources and the use of specific corpus-linguistic means for the forming of hypotheses and the validation of qualitative analyses:

> "Semantic battle" is understood here – formulated generally – as the attempt to ascertain certain linguistic forms in a knowledge domain as an expression of specific, interest-driven, and action-guiding thought patterns (Felder 2006c).

In discourse linguistics, particularly within the research network *Sprache und Wissen,* the concept of "knowledge domain" used here by Felder leads to the differentiation of the field of research that is currently being pursued, almost 30 years after the seminal contribution contribution made by Busse & Teubert. This differentiation can be seen in the still incomplete *Handbuch Sprache und Wissen* (cf. Felder & Gardt 2015), which remains the most important resource for documenting the current state of discourse linguistics in German Studies, alongside the "Werkzeuge" in *Forschen in der Linguistik* (Beißwenger et al. 2022) as well as the ongoing publication series *Sprache und Wissen*[3].

References

Beißwenger, M., Lemnitzer, L., & Müller-Spitzer, C. (Eds.). (2022). *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium.* Paderborn: Brill | Schöningh-Fink.

Braudel, F. (1977) Geschichte und Sozialwissenschaften. Die longue durée. In C. Honegger (Ed.), *Schrift und Materie der Geschichte. Vorschläge zur systematischen Aneignung historischer Prozesse* (pp. 47–85). Frankfurt a.M.: Suhrkamp.

Bubenhofer, N. (2009). *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse.* Berlin/New York: De Gruyter.

Busse, D. (1987). *Historische Semantik. Analyse eines Programms.* Stuttgart: Klett-Cotta [= *Sprache und Geschichte,* 13].

3 See https://www.degruyter.com/serial/suw-b (Accessed: 17 June 2024).

Id. (1997). Das Eigene und das Fremde. Annotationen zu Funktion und Wirkung einer diskurssemantischen Grundfigur. In M. Jung, M. Wengeler & K. Böke (Eds.), *Die Sprache des Migrationsdiskurses. Das Reden über 'Ausländer' in Medien, Politik und Alltag* (pp. 17–35). Opladen: Westdeutscher Verlag.

Id. (2000). Historische Diskurssemantik. Ein linguistischer Beitrag zur Analyse gesellschaftlichen Wissens, *Sprache und Literatur in Wissenschaft und Unterricht,* 31(86), 39–53.

Id. (2007). Diskurslinguistik als Kontextualisierung. Sprachwissenschaftliche Überlegungen zur Analyse gesellschaftlichen Wissens. In I. H. Warnke (Ed.), *Diskurslinguistik nach Foucault. Theorie und Gegenstände* (pp. 81–105). Berlin/New York: De Gruyter [= *Linguistik – Impulse & Tendenzen,* 25].

Id. (2013). Linguistische Diskurssemantik. Rückschau und Erläuterungen nach 30 Jahren. In Id. & W. Teubert (Eds.), *Linguistische Diskursanalyse. Neue Perspektiven* (pp. 31–53). Wiesbaden: Springer.

Id., & Teubert, W. (1994) Ist "Diskurs" ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik. In eid. & F. Hermanns (Eds.), *Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik* (pp. 10–28). Opladen: Westdeutscher Verlag. Erneut abgedruckt in id. & W. Teubert (Eds.). (2013). *Linguistische Diskursanalyse. Neue Perspektiven.* Wiesbaden: Springer.

Felder, E. (Ed.). (2006a). *Semantische Kämpfe. Macht und Sprache in den Wissenschaften.* Berlin/New York: De Gruyter. DOI: https://doi.org/10.1515/9783110193282 (Accessed: 17 June 2024).

Id. (2006b). Zur Intention des Bandes. In id. (Ed.), *Semantische Kämpfe. Macht und Sprache in den Wissenschaften* (pp. 1–11). Berlin/New York: De Gruyter.

Id. (2006c). Semantische Kämpfe in Wissensdomänen. Eine Einführung in Benennungs-, Bedeutungs- und Sachverhaltsfixierungs-Konkurrenzen. In id. (Ed.), *Semantische Kämpfe. Macht und Sprache in den Wissenschaften* (pp. 13–46). Berlin/New York: De Gruyter.

Id. (2008). Das Forschungsnetzwerk "Sprache und Wissen". Zielsetzung und Inhalte, *Zeitschrift für Germanistische Linguistik,* 36, 270–276.

Id., & Gardt, A. (Eds.). (2015). *Handbuch Sprache und Wissen.* Berlin/Boston: De Gruyter [= *Handbücher Sprachwissen,* 1]. DOI: https://doi.org/10.1515/9783110295979 (Accessed: 17 June 2024).

Eid. (Eds.). (2018). *Wirklichkeit oder Konstruktion? Spachtheoretische und interdisziplinäre Aspekte einer brisanten Alternative.* Berlin/Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110563436 (Accessed: 17 June 2024).

Felder, E., & Jacob, K. (2017). Diskurse. In Th. Niehr, J. Kilian & M. Wengeler (Eds.), *Handbuch Sprache und Politik* (pp. 389–406). Bremen: Hempen [= *Sprache – Politik – Gesellschaft,* 21(1)].

Felder, E., & Müller, M. (Eds.). (2009). Wissen durch Sprache. Theorie, Praxis und Erkenntnisinteresse des Forschungsnetzwerks “Sprache und Wissen”. Berlin/New York: De Gruyter.

Eid., & Vogel, F. (Eds.). (2012a). *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analyse.* Berlin/Boston: De Gruyter [= *Linguistik – Impulse & Tendenzen,* 44].

Eid. (2012b). Korpuspragmatik. Paradigma zwischen Handlung, Gesellschaft und Kognition. In eid. (Eds.), *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analyse* (pp. 3–30). Berlin/Boston: De Gruyter [= *Linguistik – Impulse & Tendenzen,* 44].

Fladrich, M. (2022). INCEpTION – ein Werkzeug für die kollaborative Annotation. In M. Beißwenger, L. Lemnitzer & C. Müller-Spitzer (Eds.), *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium* (pp. 503–515). Paderborn: Brill | Schöningh-Fink.

Foucault, M. (1978/2000). *Dispositive der Macht. Über Sexualität, Wissen und Wahrheit.* Berlin: Merve.

Id. (1981). *Archäologie des Wissens.* Übers. des franz. Originals von 1969. Frankfurt a. M.: Suhrkamp.

Gardt, A. (2017). Zum Diskursbegriff, *Der Deutschunterricht,* 69, 2–7.

Göhring, Th. (2023). *Diskursive Kämpfe. Agonalität im politischen Sprachgebrauch am Beispiel des französischen Präsidentschaftswahlkampfs.* Berlin/Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110981537 (Accessed: 17 June 2024).

Haaf, S. (2022). Werkzeuge für die Korpusanalyse. In M. Beißwenger, L. Lemnitzer & C. Müller-Spitzer (Eds.), *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium* (pp. 471–489). Paderborn: Brill | Schöningh-Fink.

Hermanns, F. (1994). Linguistische Anthropologie. Skizze eines Gegenstandsbereiches linguistischer Mentalitätsgeschichte. In D. Busse, id. & W. Teubert (Eds.), *Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik* (pp. 29–59). Opladen: Westdeutscher Verlag.

Id. (1995). Sprachgeschichte als Mentalitätsgeschichte. Überlegungen zu Sinn und Form und Gegenstand historischer Semantik. In A. Gardt, K. J. Mattheier & O. Reichmann (Eds.), *Sprachgeschichte des Neuhochdeutschen. Gegenstände, Methoden, Theorien* (pp. 69–99). Tübingen: Niemeyer.

Id. (2001). Neue deutsche Sprachgeschichte, *Mitteilungen des Deutschen Germanistenverbandes,* 48, 572–600.

Horbach, A. (2022). Werkzeuge für die automatische Sprachanalyse. In M. Beißwenger, L. Lemnitzer & C. Müller-Spitzer (Eds.), *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium* (pp. 490–502). Paderborn: Brill | Schöningh-Fink.

Jäger, S. (2015). *Kritische Diskursanalyse. Eine Einführung.* 7. ed. Münster: Unrast.

Kämper, H. (2017). Personen als Akteure. In K.S. Roth, M. Wengeler & A. Ziem (Eds.), *Handbuch Sprache in Politik und Gesellschaft* (pp. 259–279). Berlin/Boston: De Gruyter [= *Handbücher Sprachwissen,* 19].

Ead. (2022). Emotion und Sprachgebrauch. Ein linguistischer Beitrag zur Entstehungsgeschichte des Nationalsozialismus. In M. Beißwenger, L. Lemnitzer & C. Müller-Spitzer (Eds.), *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium* (pp. 122–139). Paderborn: Brill | Schöningh-Fink.

Niehr, Th. (2014). *Einführung in die linguistische Diskursanalyse.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Roth, K.S. (2015). *Diskursrealisationen. Grundlegung und methodischer Umriss einer pragmatisch-interaktionalen Diskurssemantik.* Berlin: Erich Schmidt [= *Philologische Studien und Quellen,* 247].

Id., M. Wengeler, & A. Ziem (Eds.). (2017). *Handbuch Sprache in Politik und Gesellschaft.* Berlin/Boston: De Gruyter [= *Handbücher Sprachwissen,* 19].

Schmidt, Th. (2022) Werkzeuge für die Transkription gesprochener Sprache. In M. Beißwenger, L. Lemnitzer & C. Müller-Spitzer (Eds.), *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium* (pp. 451–460). Paderborn: Brill | Schöningh-Fink.

Spieß, C., & Köpcke, M. (2015). *Metapher und Metonymie. Theoretische, methodische und empirische Zugänge.* Berlin/Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110369120 (Accessed 17 June 2024).

Spitzmüller, J., & Warnke, I.H. (2011). *Diskurslinguistik. Eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse.* Berlin/New York: De Gruyter.

Stefanowitsch, A. (2020). *Corpus linguistics. A guide to the methodology.* Berlin: Language Science Press. [= *Textbooks in Language Sciences,* 7]. DOI: https://doi.org/10.5281/zenodo.3735822 (Accessed: 17 June 2024).

Id. & Gries, S.T. (2009). Corpora and Grammar. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics* (pp. 933–952). Berlin: Mouton de Gruyter [= *Handbooks of Linguistics and Communication Science HSK,* 29(1)].

Teubert, W. (2006). Korpuslinguistik, Hermeneutik und die soziale Konstruktion der Wirklichkeit, *Linguistik Online,* 28, 41–60. URL: http://www.linguistik-online.de/28 (Accessed: 17 June 2024).

Warnke, I.H. (Ed.). (2007). *Diskurslinguistik nach Foucault. Theorie und Gegenstände.* Berlin/Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110920390 (Accessed: 17 June 2024).

Wengeler, M. (2003). *Topos und Diskurs. Begründung einer argumentationsanalytischen Methode und ihre Anwendung auf den Migrationsdiskurs (1960–1985).* Tübingen: Niemeyer [= *Reihe Germanistische Linguistik,* 244].

Id. (2017). Wortschatz I. Schlagwörter, politische Leitvokabeln und der Streit um Worte. In K.S. Roth, M. Wengeler & A. Ziem (Eds.), *Handbuch Sprache in Politik und Gesellschaft* (pp. 22–46). Berlin/Boston: De Gruyter [= *Handbücher Sprachwissen,* 19].

Wolfer, S, & Hansen, S. (2022). Werkzeuge für die statistische Analyse. In M. Beißwenger, L. Lemnitzer & C. Müller-Spitzer (Eds.), *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium* (pp. 461–479). Paderborn: Brill | Schöningh-Fink.

Figure Credit

Fig. 1: Model of Discourse Linguistic Multilevel Analysis (DIMEAN) according to Spitzmüller & Warnke 2011, 201.

Named Entity Recognition

Evelyn Gius

https://orcid.org/0000-0001-8888-8419

Abstract This chapter introduces the automatic recognition of entities in texts using the method of *Named Entity Recognition.* After defining *Named Entities,* initial considerations regarding their recognition are presented. Then, the chapter outlines the development of *Named Entity* systems in language processing and the most important associated models. The applicability of *Named Entity Recognition* in theology is then examined and practical tips for testing *Named Entity* systems are provided. The chapter concludes with references to tools and resources for *Named Entity Recognition.**

Keywords Named Entity Recognition, Entities, Proper Names, Language Processing

1. What is *Named Entity Recognition*?

In language processing, words or expressions in a text referring to specific entities in the world are referred to as *Named Entities* and their automatic recognition as *Named Entity Recognition.*[1] *Named Entities* specifically include expressions for specific persons, places, or organizations. In principle, *Named Entities* have a clearly defined identity and can be identified by a name or a specific term, i.e., they can be named. In addition to proper names, *Named Entities* also include other designations. For example, both the personal proper name "Hildegard," and the specific term "The Master of Rupertsberg," are *Named Entities* of the personal entity type.

The recognition of *Named Entities,* also known as *Named Entity Recognition* (NER), is an important method in text processing and analysis. NER systems identify entities and classify them into predefined categories such as "persons," "places," or "organizations." To do so, the systems use machine learning methods based on linguistic information and semantic correlations in texts. Some systems also use lists of known proper names, which are called *Gazetteers* in language processing. These are particularly helpful for places and other geopolitical entities, whereby the recognition results – possibly counterintuitively – are better when using fewer, highly

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

1 For many disciplines in the humanities, one must of course add: in addition to real world entities, these entities can also be present in narrated worlds.

Evelyn Gius: Named Entity Recognition, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 229–243. DOI: https://doi.org/10.11588/heibooks.1521.c21951

frequent proper names. Extensive lists of less frequent proper names, on the other hand, diminish the results.

Like all computational methods in language processing, NER systems were initially developed based on rules, whereas nowadays machine learning methods achieve better results, although the phenomena previously described in the rules do play a role (see the section “The development of NER systems”). Regardless of the technology used, *Named Entity Recognition* systems use sequence tagging approaches, in which each element of a sequence is assigned a value. For example, each word in a text is assigned the information as to whether it is a *Named Entity* and, if so, which class. Strictly speaking, *Named Entity Recognition* consists of two tasks: Recognizing *Named Entities* (identification) and classifying the recognized *Named Entities* into the predefined classes (classification).

The *Named Entity* classes used differ depending on the system. Many NER systems recognize the classes persons, locations, and organization (Tab. 1), which are typically designated as PER (cf. *person*), LOC (cf. *location*), ORG (cf. *organization*). Most systems also have a fourth class, comprising either of geopolitical entities (GPE, for *geo-political entity*) or a residual class (MISC, for *miscellaneous*). In addition to persons, places, and organizations, other classes and corresponding expressions are also considered *Named Entities,* e.g., dates (“September 17, 1179”), quantities (“five kilograms”), abstract terms (“religion”) and general classes (“monastery”).

Tab. 1 The three most common *named entity* classes in NER systems.

Class	Tag	Example for Entities	Example (*Named Entity* in bold)
Person	PER	human, figures, saints	**Abigajil** prevents further violence.
Location	LOC	cities, mountains, countries, bodies of water	There are numerous regions in **South Asia.**
Organization	ORG	companies, associations, institutions	The **Roman Catholic Church** is the largest Christian church.

Understood within the context of language processing, *Named Entity Recognition* is a well-established and widely used technique. Along with other basic operations – such as the segmentation of text into word and sentence units *(Tokenization, Sentence Splitting)* and the tagging of word types and syntactic units *(Part-of-Speech Tagging, Dependency Parsing)* – NER is a pre-processing step in most language processing pipelines.[2] In machine learning processes, *Named Entities* are used as a *feature,* i.e., an aspect

2 On the structure of language processing pipelines, see Biemann et al. (2022, 85 ff.). The introduction is also suitable for deepening some of the other language processing methods mentioned here.

included in the analysis of texts in a wide variety of tasks, whereby the systems calculate an appropriate consideration (the so-called weighting) of the *feature* in the course of the learning process.[3]

In addition to the recognition of *Named Entities* in texts, typical applications of NER in language processing include methods that are based on the results and analyze further semantic information. These methods include the extraction of relationships between the entities (e.g., family relationships between persons, spatial relationships between places or persons and places, etc.), the creation of knowledge graphs in databases used for search engines, and the recognition of events, enabling further semantic textual analysis. The fields in which NER is used are correspondingly diverse. They range from scientific research to government institutions and companies. NER is also used to create market analyses, track customer feedback, and gain intelligence on potential threats, as well as to analyze historical texts and examine cultural developments.

2. An Initial Approach to *Entity Recognition*

A few examples of possible textual features that could be used to identify *Named Entities* systematically with suffice for illustrating the approach.

> Example 1: In those days, a decree went out from **Caesar Augustus** that all the world should be registered. This registration was the very first and took place at the time when **Quirinius** was governor of **Syria.**

In this example, two personal entities ("Caesar Augustus" and "Quirinus") and one location entity ("Syria") are mentioned. Textual features used to recognize these entities could be the spelling. In German, as some other languages, proper names are capitalized.[4] Another characteristic of personal entities is also that proper names do not normally have an article, distinguishing them from other nouns. The term "Caesar Augustus" also includes the title "Caesar." One can thus formulate a rule that titles and subsequent capitalized words denote personal nouns. "Syria" is also a proper noun recognizable by its capitalization. Moreover, certain prepositions such as "in," "of," etc. can refer to a location entity.

3 On the use of features in machine learning methods, see Jurafsky & Martin (2023, 59; 60ff.). The introduction is also suitable for in-depth study of *Named Entity Recognition* and all other language processing methods mentioned.

4 In German, however, capitalization applies not only to proper names, but also to nouns, which means that many other words also have this feature and makes it less easy to recognize *Named Entities* than in English, for example.

Example 2: **Saint Teresa of Ávila** was born in **Spain** in the **16th century.** Her mystical experiences led to important writings.

In this example, there are three mentioned entities: a person entity ("St. Teresa of Ávila"), a date entity ("16th century"), and a location entity ("Spain"). The mentioned features can be used for the person and place entities. The "saint" is also a kind of title, although one must conjure up rules for saints, such as the combination of the adjective "saint" preceded by "of" and a place name. For the date, one might define a series of formats that typically combine numbers, punctuation marks, and words, thus distinguishing them from other expressions.

Example 3: **Francis of Assisi** founded the **Franciscan order** in **Italy.**

The bold words in the third example are: a person entity ("Francis of Assisi"), an organization entity ("Franciscan order"), and a place entity ("Italy"). For "Francis of Assisi," a partial rule of the rule of saints could be used, particularly, the scheme [proper name] "of" [place entity]. The same rules apply to "Italy" as to "Syria" and "Spain" in the previous examples. The "Franciscan Order," on the other hand, can be recognized by the fact that the expression consists of a capitalized but not very frequent word introduced with a definite article. Presumably, a rule can also be derived from the composition, since "Franciscan" is a name derived from a proper name and "order" is a general organizational term.

The considerations on the three examples are intended to show that *Named Entities* can be distinguished from other expressions based on textual features. These features include spelling, the use of certain prepositions, or other combinations of word type sequences, as well as features on the character level such as capitalization, letter sequences, or the use of characters atypical for other word types such as numerals or punctuation, the syntactic structure (where in the sentence might one expect *Named Entities?*), or even typical contexts or occurrence frequencies of *Named Entities.* These features were initially used in the recognition of *Named Entities* based on corresponding rules.

3. The Development of NER Systems

The emergence of *Named Entity Recognition* goes back to the beginnings of computational processing of natural language in the 1950s and 60s. During this time-period, word processing systems were developed for analyzing basic linguistic information. Overall, the history of NER corresponds to the development of many language processing applications, ranging from rule-based recognition of phenomena to machine learning methods and *Deep Learning* approaches.

The first NER approaches focused mainly on identifying the names of people and places. For these identifications, they defined rules or patterns to target specific properties of proper names as in the examples discussed above. These rule-based methods enabled the identification of names in texts based on certain characteristics such as capitalization or special characters. However, heuristic approaches are limited. They did not achieve satisfactory results due to the variety of named entities and contexts in which they occur.

The use of machine learning techniques in NER, which emerged in the 1990s, led to an improvement in the systems.[5] Statistical models and machine learning algorithms were used to recognize and classify named entities based on previously manually annotated training data. *Hidden Markov* models (HMM) and *Maximum Entropy* models were used, which can take context information and statistical probabilities into account in the NER. *Hidden Markov* models can analyze the sequence of words in a text and calculate the probability of a word being a *Named Entity.* They assume hidden states unknown at the beginning (in the case of NER: the entities), as well as observable states consisting of the words in the text. The model is trained to optimize the transition probabilities between these states and the output probabilities for each word. *Entity Recognition* is based on the most probable state transitions determined in this way, which establish a link between the unknown (or hidden) entity classes and the observable words. *Maximum Entropy* models (MaxEnt) are also probabilistic models. They are based on maximum entropy principles that optimize probabilities for a set of classes or categories. In NER, maximum entropy models predict the association of *Named Entity* categories by using trained weightings of appropriate text features. These features could be words, contextual information, capitalization, etc. The aim is to adjust the weights of the features so that the probability for each entity category is calculated in terms of maximum entropy.

The performance of NER systems has been further increased by the establishment of *Deep Learning.* Artificial neural networks, which "learn" phenomena in a large amount of data using several layers, are now being used to recognize *Named Entities.* The first *Deep Learning* systems to be used were *recurrent neural networks* (RNNs), which have been around since the 1980s but were only used in NER in the 2000s. RNNs are neural networks that have been specifically developed for processing sequential data. They can be used to process the sequence of words in a text and calculate the probability of each word belonging to a particular class. Unlike the models mentioned above (HMMs and MaxEnt), the RNN also considers the context of the previous words, thus enabling a more precise recognition of entities. However, RNNs have difficulties in processing long sequences. The next development, *Long Short-Term Memory* networks (LSTM), then included the capturing of long-term dependencies in sequences. In the NER, LSTMs enable more precise modeling of relationships

5 For a brief overview of the systems developed from the early stages to the current *Transformer*-based approaches, cf. Jurafsky & Martin (2023, 183).

between words and the recognition of entities that can vary over longer sections. LSTMs can effectively utilize both local and global contextual information, advancing the capabilities of NER. A further improvement of NER systems in the early 2010s, combined bi-directional LSTM models with *Conditional Random Fields* (CRFs). Bi-directional LSTMs not only capture the context of the sequence before a particular word, but also the context after the word, which increases the quality of NER. The use of CRFs helps to model dependencies between adjacent words and their classification, which enables a more coherent assignment of entity labels. The recognition of nested entities has also been significantly improved by *Deep Learning* methods.

The current *State of the Art* systems for NER are based on pre-trained *Transformer* models such as BERT or GPT, which have been under development since the mid-2010s. *Transformers* are a further development of recurrent networks in which context-dependent information can be calculated simultaneously through *self-attention* and *memory* layers. These networks can therefore be trained on a large amount of text data and generate even better language models. Using *Transfer Learning*, these general models can then be adapted for specialized tasks such as NER.

Regardless of the models, BIO annotation (Ramshaw & Marcus 1995) for NER has become the standard approach for sequence labeling in a span recognition problem. The approach provides three labels that also account for the boundaries of the *Named Entities.* This method captures each word (or *token*) of a *Named Entity* expression as follows: The first word is given the label B (for *begin*), all following words are labeled I (for *inside*) and all words outside the *Named Entity* are labeled O (for *outside*). There are separate B and I labels for each entity class to map these. For the beginning of example 1, a BIO annotation would look like this:

The	Saint	Teresa	of	Avila	was	in	…
O	B-PER	I-PER	I-PER	I-PER	O	O	

Fig. 1 Sequence encoding of a *Named Entity* of Person (PER) with BIO labels

4. Challenges of Automatic *Entity Recognition*

While NER is one of the basic methods of language processing and the current NER systems achieve good results, there are still some persistent challenges in the recognition of *Named Entities.*

The linguistic form of *Named Entities* is very diverse. The wide range of inflections, derivations, morphs or syntactic rules, and word order in a language increase the complexity of recognition. NER is thus particularly difficult in morphological languages such as Hebrew. There is also a practical problem: NER systems are based on extensive training data, so the performance of the systems depends on the availability of sufficient suitable data in the relevant language. The development of universally applicable NER systems is even more difficult by the linguistically and culturally dependent differences in grammar, syntax, and nomenclature, i.e., the way in which entities are named.

Named Entities are not only multiword phrases – such as “University of Tübingen” or “Mary Magdalene” – they are also sometimes nested. For the correct recognition of entities such as “Apostle Paul” or “Hildegard of Bingen,” the words belonging to the proper name must not only be recognized as the title (“Apostle” or “Bingen”); they must also be determined as belonging to the personal entity, which requires deeper semantic processing and better modeling of contexts in the text.

The fact that *Named Entities* can also be multiword phrases also makes the evaluation of NER systems more complex than with uniform segments. In contrast to *Part of Speech Tagging,* for example, where a value is assigned to each individual word, or to classification tasks that are performed for entire texts, the textual span that the respective entity covers must be determined for the NER. Given that words are typically the training unit for the NER and the output unit is entities – potentially multiword expressions – there is an incongruity. Accordingly, in the BIO annotation system, only partially recognized multiword entities are evaluated incorrectly several times because the annotations are incorrect due to the missing words (the B annotation comes one or more words too late or the O annotation too early, with corresponding consequences for the I annotations). This problem concerns the non-recognition of the same entity being evaluated in the same way and thus considered equally good or bad. However, this problem can be mitigated by a corresponding error weighting in the evaluation.

Languages considered to be data-poor, e.g., pre-modern languages, present a particular evaluative problem. There is often no further annotated data in these types of languages that can be used as *benchmarks* to check whether the evaluated NER system also achieves similarly good results with unknown texts or whether there is an *overfitting* on the training data, where only these instances are recognized.

There are several challenges that are more prerequisites. For example, a NER system only recognizes the entities in the text that are relevant to a question if they are named explicitly and with clearly defined names or expressions. NER is therefore not suitable for recognizing pronouns, generic expressions, unspecific terms, and indirect references to *Named Entities.* Moreover, there are difficulties in recognizing entities like abstract concepts and entities if named by infrequently used technical terms or local names.

While the latter difficulties are certainly addressed by NER systems, the recognition of pronouns or the like is not part of NER, for which are many pragmatic reasons. One reason is that the additional challenges of the closely related task of coreference resolution would also have to be solved. Coreference resolution is determining when pronouns, demonstrative expressions, or other referential elements in the text refer to previously mentioned entities. Coreference resolution requires an in-depth understanding of the context and semantic relationships of a text. Additionally – esp. in the field of theology – there are relevant questions of identity, since all expressions referring to the same entity must be identified. This problem is often difficult in less obvious cases than the Trinity because the identity of many entities is difficult to ascertain, e.g., in the case of temporal or other changes. For example, a school of thought can be perceived as the same over decades or divided into certain sections, several organizational entities, such as a family, can be a single entity or the addition and removal of family members can each be perceived as new families, or the life phases of a person with very different views and actions can also be perceived as separate personal entities.[6]

5. NER in Theology?

Since the techniques of automatic language processing can be used in any science focusing on text analysis, NER can be used in theological research.[7] In principle, applying these methods is possible and useful in all areas in which entities such as persons, places, dates, or concepts or their relationship to each other are relevant to a research interest. Potential fields of application range from the identification of specific phenomena in individual texts to the analysis of large volumes of text or corpora. In addition to identifying the corresponding *Named Entities,* the NER is also suitable for analyzing the distribution, interrelationship, and developments over time, of the entities. Developments can also be compared with different text groups or grouping of texts based on *Named Entities.* An analysis based on *Named Entities* can be aimed at the question of the most frequent mentions of actors or places in religious texts or the quantitative comparison of the respective proportions of mentions between different texts or text groups. Questions about the first mention and subsequent development of the frequency of mentions of persons, locations, or concepts in a corpus of diachronic texts, i.e., texts that cover a longer period, can also be analyzed. An NER can be used to carry out stylistic analyses – such as in homiletics – or to identify texts in a corpus that relate to a specific topic recognizable via *Named Entities.*

These types of applications lead to interesting findings with the NER relevant for theology. Nevertheless, *Named Entity Recognition* is not widespread in theology and has not yet had any recognizable significance in publications relevant to digital

6 For an in-depth consideration, see the *Stanford Encyclopedia of Philosophy.* On the identity problem, see Noonan & Curtis (2022). On the problem of fictional entities, see Kroon & Voltolini (2023).

7 For an overview of language processing methods in the humanities, see Piotrowski (2012); Sporleder (2010); and the methodological introductions in the forTEXT portal at https://fortext.net/routinen/methoden (Accessed: 17 June 2024).

approaches.[8] There are several reasons why NER has not gained any traction in religious studies. One reason is that the application of language processing techniques in the humanities is generally still a relatively young branch of research beyond computer and corpus linguistics. In addition, there is a hesitancy towards the use of computational tools in theology as well as in other humanities that work more exemplarily or hermeneutically. Finally, the so-called operationalization of a question, i.e., the translation of the question into qualities that can be measured by *Named Entities,* is not a trivial task, which is methodologically contrary to the established practices of theological text analysis. However, if recent developments in the field of *Digital Theology* are examined, one can assume that some progress will also be made in the field of Computational Theology in the coming years and that NER methods will also be used. However, even if any reservations have been dispelled and the necessary skills for the implementation of NER are available, there are limitations to the quality of the analyses that must be considered.

6. Notes on the use of NER systems

Like most language processing methods, NER systems are typically developed for English and based on news articles or texts found on the internet. Therefore, for languages other than English, or for text types other than news and internet texts, the quality of available systems decreases. Moreover, the results are often different depending on the *Named Entity* class. While the classic categories for persons, locations, and organizations are usually well recognized and achieve recognition rates of over 90% in the better systems, the recognition quality for other categories is considerably lower. Nonetheless, the NER can also be used in cases where the systems do not work optimally if prepared and implemented accordingly.

Prior to using a NER system, one should assess the extent to which the quality of the recognition is sufficient to make reliable statements based on the results. In language processing, results with a F1 value of 0.8 or more are considered very good, while results of 0.95, which are now achieved in NER for English – and occasionally

8 For example, NER is only mentioned once in Heyden & Schröder (2020) or Sutinen & Cooper (2021). NER is not mentioned at all in the publication series *Introductions to Digital Humanities – Religion* (ed. by Claire Clivaz, Frederik Elwert, Kristian Petersen, Ortal-Paz Saar and Jeri Wieringa) nor in the *Digital Biblical Studies* (ed. by Claire Clivaz and Ken M. Penner). Searches in catalogs were also virtually fruitless: a search for NER in the Religious Studies Bibliography of the Specialized Information Service (FID) at https://www.relbib.de (Accessed: 17 June 2024) yields only one hit (Blouin 2021), which is potentially relevant but not pertinent. There are no theological titles among the results of the search in the University of Frankfurt catalog for “Named Entity Recognition.” Even if there are individual publications not found in these search attempts, the lack of results indicates at least a low relevance of NER in theology to date.

for other languages such as German – are considered (almost) perfect. The F1 value is made up of the values for the measures of recall and precision. Accordingly, an F1 value of 0.8 means that the average proportion of phenomena found in the text (recall) and of correctly identified passages among the passages found (precision) is 80%. Since the value is an average of the two values and these are in turn calculated for several subcategories (persons, locations, organizations, etc.), the F1 value does not indicate anything about the quality of recognition for specific aspects. The F1 value is therefore – like any evaluation measure – only a guide value for the actual quality of the application. The value usually indicates nothing about the suitability of the system for the specific research interest. One must first check the extent to which a system delivers suitable results for the research question and the text corpus used. A quality check is more important if further steps based on the NER are implemented automatically, such as the recognition of entity relations or the coreference resolution, in which all entity names and other possible references – such as pronouns – to one and the same entity are recognized.

If a NER system is used for texts that differ from the texts used and evaluated during the development of the system, a specific check of the recognition quality should therefore be carried out beforehand. Ideally, the system should be evaluated based on an annotated test data from the corpus used, i.e., a meaningful F1 value should be created for the specific research requirements. However, at least a sample check of the output results and individual text parts should be carried out regarding the phenomena found. The sample check can be used to assess whether a system correctly recognizes the phenomena searched for and to what extent it may include false phenomena. In addition, possible systematic errors can be recognized, e.g., whether a location name is incorrectly recognized as a personal name, whether certain multiword expressions are not or only partially recognized, or whether individual terms tend not to be recognized. Such errors can greatly distort the text analysis, depending on the type of error. For example, if one wishes to compare the relevance of certain concepts in texts, one should ensure that one of the concepts is not recognized significantly worse than the others and is therefore found less frequently in the texts.

If the quality of the system is unsatisfactory and cannot be used for automatic analysis, there are still two ways for it to be used, both involving a further manual check of the results and thus ensure an analysis based on them. First, each system can be used as a heuristic system and point out any interesting aspects of the analyzed texts. Even if the results of the NER are not evaluated quantitatively – which should not be done anyway if the results are not good enough – their results can be used as an indication of potentially interesting texts or text passages. Perhaps the NER can be used to find a text that has not previously been recognized as relevant in a certain context, or one comes across terms that have not yet been considered, although they were prominent and relevant at a certain time or in certain texts. The results may also reveal connections due to the common occurrence of entities that have not previously been considered.

Second, NER systems not suitable for automatic analysis can be used as a pre-processing step that provides data for subsequent manual processing. If a system has an acceptable *recall,* i.e., finds a good proportion of the phenomena searched for, the quality of the data can be improved significantly by manual processing. To do so, the incorrect results are sorted out. The remaining data can then be used in further – even manually supported – steps of analysis or for a quantitative evaluation. This approach is viable for coreference resolution in longer texts because checking and correcting the coreference chains requires comparatively little effort. Manual processing essentially consists of correcting the incorrect mentions of entities in the coreference chains, which contain all mentions of an entity in a text, and rejoining chains that may have been separated due to recognition errors. Depending on the knowledge gained from the data prepared, manual checks are to be considered.

7. Tools and Resources

There have been many types of NER systems developed in recent decades. When selecting systems and platforms, one should first note that rule-based NER methods are often suitable for simpler cases, while more complex scenarios may require machine learning. In addition, ideally several systems should be tested on the same data to identify the most suitable system. Currently, three *open source* systems are widely used in applications: the *Natural Language Toolkit* (NLTK),[9] *spaCy,*[10] and the *Stanford Named Entity Recognizer.*[11] All three achieve good results for various natural languages and are regularly updated. Their *Python* or *Java*-based models are comparatively easy to use. However, one might also benefit from a search for other language specific NER systems.[12] Platforms that allow the assembly of one's own processing pipeline are also particularly interesting for users who are not (yet) experienced. The German platform *WebLicht* is freely accessible to members of many scientific institutions and pro-

9 See https://www.nltk.org. For the use of NER, see https://www.nltk.org/book/ch07.html. All addresses mentioned in this section were accessed on 17 June 2024.

10 See https://spacy.io/models. For the use of NER, see https://spacy.io/universe/project/video-spacys-ner-model-alt.

11 See https://stanfordnlp.github.io/CoreNLP/ner.html. For the use of the Pipeline, see https://stanfordnlp.github.io/CoreNLP/pipeline.html.

12 There are good approaches for Latin (see, e.g., Erdmann et al. 2016), Ancient Greek (see, e.g., Yousef et al. 2022), Hebrew (see, e.g., Bareket & Tsarfaty 2021), and premodern or classical languages (see, e.g., Johnson et al. 2021 and Burns 2019).

vides various systems for both pre-processing and NER itself, which can be combined on a graphical interface and applied to provided texts in many languages.[13]

To develop NER systems, one must first select suitable data. There are several annotated corpora that can be reused depending on the field of application, such as the English corpus for literary texts by Bamman et al. (2019), or the German newspaper text corpora (among others) by Tjong Kim Sang & De Meulder (2003) and Benikova et al. (2014). Further annotation of pre-processed data for the NER may be particularly useful for languages with fewer resources (e.g., for Latin in the *EvaLatin* corpus by Sprugnoli et al. (2020), which is already enriched with information on lemmatization and *Part of Speech Tagging*).

Existing directories can often be reused for the creation of *gazetteers.* In principle, large directories, ideally freely available under appropriate licenses like the Creative Commons license, are suitable for this purpose. For example, corresponding *Wikipedia* categories can be used (such as man, woman, figure, saint for personal names or corresponding categories for locations, etc.) to obtain entity names.[14] Another source is the *Gemeinsame Normdatei* (GND), which provides authority data from catalog data on persons and other areas in a range of metadata and data services, which is also worth searching for specific data.[15] For historical texts, such as the ruling class of the Roman Empire in the early and high imperial period, the encyclopedia of persons of the Berlin-Brandenburg Academy of Sciences and Humanities might be used,[16] or even the lexicon of Greek personal names of the University of Oxford.[17] Institutions like the EU or individual states also provide numerous data relevant to the NER. The EU offers a large directory of names as well as a range of other information,[18] and the U.S. Geological Survey (USGS) provides various data on locations and other geological information.[19] Many internet directories are available.

There are two introductory texts recommended, one in German and one in English – the German-language exercise by Schumacher (2019) on adapting the *Stanford Named Entity Recognizer* for literary texts, which is also suitable for beginners, and the English-language introduction by Grunewald et al. (2022), which provides a low-threshold introduction to a *Python* analysis of locations in data on prisoners of war and explains how to integrate a *gazetteer.*

13 See https://weblicht.sfs.uni-tuebingen.de/weblicht/. For a description of the available NER models, see https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tools_in_Detail#Named_Entity_Recognition.
14 To perform these tasks structurally, among others, see https://www.wikidata.org.
15 Cf. the GND service *Entity Facts* at https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/Entity-Facts/entityFacts_node.html.
16 Cf. *Prosopographia Imperii Romani saec. I. II. III.,* available at https://pir.bbaw.de.
17 Cf. https://www.lgpn.ox.ac.uk.
18 For an overview, cf. https://data.jrc.ec.europa.eu, and for a list of names, see https://data.jrc.ec.europa.eu/dataset/jrc-emm-jrc-names.
19 Cf. https://www.usgs.gov/products/data/all-data.

References

Bamman, D., Popat, S., & Shen, Sh. (2019). An Annotated Dataset of Literary Entities. In *Proceedings of the 2019 Conference of the North* (pp. 2138–2144). Minneapolis, Minnesota: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/N19-1220 (Accessed: 17 June 2024).

Bareket, D., & Tsarfaty, R. (2021). Neural Modeling for Named Entities and Morphology (NEMO2), *Transactions of the Association for Computational Linguistics,* 9, 909–928. DOI: https://doi.org/10.1162/tacl_a_00404 (Accessed: 17 June 2024).

Benikova, D., Biemann, Ch., & Reznicek, M. (2014). NoSta-D Named Entity Annotation for German. Guidelines and Dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (pp. 2524–2531). Reykjavik: European Language Resources Association. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/276_Paper.pdf (Accessed: 17 June 2024).

Biemann, Ch., Heyer, G., & Quasthoff, U. (2022). *Wissensrohstoff Text. Eine Einführung in das Text Mining,* 2. Wesentlich überarbeitete Auflage. Lehrbuch. Wiesbaden [Heidelberg]: Springer Vieweg. DOI: https://doi.org/10.1007/978-3-658-35969-0 (Accessed: 17 June 2024).

Blouin, B., Magistry, P., & Van Den Bosch, N. (2021). Creating Biographical Networks from Chinese and English Wikipedia, *Journal of Historical Network Research,* 5(1), 303–317. DOI: https://doi.org/10.25517/JHNR.V5I1.120 (Accessed: 17 June 2024).

Burns, P.J. (2019). Building a Text Analysis Pipeline for Classical Languages. In M. Berti (Ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution* (pp. 159–176). Berlin/Boston: De Gruyter Saur [= *Age of Access? Grundfragen der Informationsgesellschaft,* 10]. DOI: https://doi.org/10.1515/9783110599572-010 (Accessed: 17 June 2024).

Ehrmann, M., Hamdi, A., Pontes, E.L., Romanello, M., & Doucet, A. (2023). Named Entity Recognition and Classification in Historical Documents. A Survey, *ACM Computing Surveys,* 56(2), 1–47. DOI: https://doi.org/10.1145/3604931 (Accessed: 17 June 2024).

Grunewald, S., & Janco, A. (2022). Finding Places in Text with the World Historical Gazetteer, *Programming Historian,* 11, no. pag. DOI: https://doi.org/10.46430/phen0096 (Accessed: 17 June 2024).

Heyden, K., & Schröder, B. (Eds.) (2020), *Theologie Im Digitalen Raum,* Gütersloh: Gütersloher Verlagshaus [= *Verkündigung und Forschung,* 65(2)].

Johnson, K.P., Burns, P.J., Stewart, J., Cook, T., Besnier, C., & Mattingly, W.J.B. (2021). The Classical Language Toolkit. An NLP Framework for Pre-Modern Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. System Demonstrations* (pp. 20–29). Online: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2021.acl-demo.3 (Accessed: 17 June 2024).

Jurafsky, D., & Martin, J.H. (2023). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 3rd ed. [Draft]. URL: https://web.stanford.edu/~jurafsky/slp3 (Accessed: 17 June 2024).

Kroon, F., & Voltolini, A. (2023). Fictional Entities. In E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy.* Stanford University: Metaphysics Research Lab. URL: https://plato.stanford.edu/archives/fall2023/entries/fictional-entities (Accessed: 17 June 2024).

Noonan, H., & Curtis, B. (2022). Identity. In E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy.* Stanford University: Metaphysics Research Lab. URL: https://plato.stanford.edu/archives/fall2022/entries/identity (Accessed: 17 June 2024).

Piotrowski, M. (2012). NLP Tools for Historical Languages. In id. (Ed.), *Natural Language Processing for Historical Texts* (pp. 85–100). Cham: Springer International Publishing [= *Synthesis Lectures on Human Language Technologies*]. DOI: https://doi.org/10.1007/978-3-031-02146-6_7 (Accessed: 17 June 2024).

Ramshaw, L., & Marcus, M. (1995). Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora.* URL: https://aclanthology.org/W95-0107 (Accessed: 17 June 2024).

Schumacher, M. (2019). Named Entity Recognition mit dem Stanford Named Entity Recognizer, *forTEXT. Literatur digital erforschen,* 1–53. URL: https://fortext.net/routinen/lerneinheiten/named-entity-recognition-mit-dem-stanford-named-entity-recognizer (Accessed: 17 June 2024).

Sporleder, C. (2010). Natural Language Processing for Cultural Heritage Domains, *Language and Linguistics Compass,* 4(9), 750–768. DOI: https://doi.org/10.1111/j.1749-818X.2010.00230.x (Accessed: 17 June 2024).

Sprugnoli, R., Passarotti, M., Cecchini, F.M., & Pellegrini, M. (2020). Overview of the EvaLatin 2020 Evaluation Campaign. In *Proceedings of LT4HALA 2020. 1st Workshop on Language Technologies for Historical and Ancient Languages* (pp. 105–110). Marseille: European Language Resources Association (ELRA). URL: https://aclanthology.org/2020.lt4hala-1.16 (Accessed: 17 June 2024).

Sutinen, E., & Cooper, A.-P. (2021). *Digital Theology. A Computer Science Perspective.* Bingley: Emerald Publishing Limited. DOI: https://doi.org/10.1108/9781839825347 (Accessed: 17 June 2024).

Tjong Kim Sang, E.F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task. Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL,* 4, 142–147. Edmonton: Association for Computational Linguistics. DOI: https://doi.org/10.3115/1119176.1119195 (Accessed: 17 June 2024).

Yousef, T., Palladino, Ch., & Jänicke, S. (2023). Transformer-Based Named Entity Recognition for Ancient Greek. In W. Scholger, G. Vogeler, T. Tasovac, A. Baillot & P. Helling (Eds.), *Digital Humanities 2023. Collaboration as Opportunity (DH2023)* (pp. 1–3). Graz: Zenodo. DOI: https://doi.org/10.5281/zenodo.8107629 (Accessed: 17 June 2024).

Topic Modeling

Melanie Althage

https://orcid.org/0000-0001-5233-1061

Abstract Topic Modeling is a method used in the Digital Humanities to examine the thematic structure of large collections of texts. This chapter offers an introduction to its methodological foundations. In addition to an overview of various Topic Modeling algorithms and their respective fields of application, the article focuses on central workflow aspects, such as the preparation of the text data (preprocessing) and evaluation of the modeling results. The aim is to provide a solid basis for the critically reflected use of Topic Modeling in theological research.*

Keywords Topic Modeling, Text Mining, Quantitative Text Analysis, Machine Learning, Natural Language Processing, Blended Reading, Distant Reading

1. Introduction

In a 2006 article, Gregory Crane asks: "What do you do with a million books" (Crane 2006)? This question has become increasingly relevant with the readily available number of digital sources (see also Stulpe & Lemke 2016, 18). However, a significant portion of these sources is only weakly structured, making it difficult to retrieve the information they contain. In what ways can this wealth of information and potential knowledge be effectively explored and made useful for research purposes? One answer to this question is provided by Topic Modeling. Topic Modeling is a clustering algorithm that uses linguistic patterns to structure large text corpora thematically and make them searchable. If one assumes that themes or content-related concepts are expressed through a specific set of terms that frequently co-occur in different historical sources, such automated pattern recognition methods can provide valuable contributions to research.

In the Digital Humanities and historical sciences, Topic Modeling has established itself as a versatile tool for a wide range of research questions. The method enables the analysis of research trends in academic journals (Mimno 2012; Wehrheim 2019; Wehrheim et al. 2022), the investigation of discourse structures in various publications

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

Melanie Althage: Topic Modeling, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 245–266. DOI: https://doi.org/10.11588/heibooks.1521.c21952

(Völkl et al. 2022; Bunout & von Lange 2019) or the positioning of Digital Humanities as a discipline in comparison to other fields (Luhmann & Burghardt 2021). In the context of theology, the method is also increasingly being used. Christopher A. Nunn, for example, presented Topic Modeling in his study as part of a broader distant reading approach and used the *DARIAH-DE Topics Explorer* (Simmler et al. 2019), a user-friendly software, to shed light on ethical topics in the letters of Augustine of Hippo (Nunn 2022). Mark Graves examined the model-theoretical and mathematical-computational aspects of Topic Modeling for his study on the moral theology of Thomas Aquinas. He demonstrated how the method can be used to analyze complex moral and theological concepts in their various facets and subsequently investigate their influence on papal encyclicals (Graves 2022).

To encourage further studies in theology, this article aims to provide a critical and reflective introduction to the method and workflow of Topic Modeling as well as its many variants and configuration possibilities. It outlines not only the potentials, but also the limitations and challenges that need to be considered when using this method in the research process. The article first outlines the basic concept of Topic Modeling. Then, the article provides an overview of various algorithms and their usage. A detailed description of the mathematical principles behind the individual methods is deliberately omitted; for in-depth information, please consult the relevant specialist literature. Finally, the article discusses the central aspects of data preparation and evaluation of the modeling results. The aim is to provide a foundation and initial orientation for the application of Topic Modeling in theological research.[1]

1 The exemplary topics presented in Figures 1–3 are derived from the German-language book reviews published on the specialist communication portal H-Soz-Kult (https://www.hsozkult.de/, accessed: 19 July 2024) between 1996 and June 2019 (15,103 reviews with approximately 18 million words). The selected topics are taken from a model comprising a total of 80 topics, created using the *Latent Dirichlet Allocation* (LDA; Blei et al. 2003) algorithm implemented in the *MALLET* software (McCallum 2002) via the Python wrapper in *Gensim* (Řehůřek & Sojka 2010) as part of the author's ongoing dissertation project, which is provisionally titled: "Mining the Historian's Web – A Method-critical Reflection on Quantitative Methods for the Analysis of Born-Digital Sources Using the Example of Historical Specialist Communication". They are based on an earlier phase of the project. The topics in Table 1, which are used as examples for illustration, are in turn based on selected German-language funeral sermons from the 17th century (299 with approximately 3 million words). These sources were digitized as part of the *German Research Foundation* (DFG)-funded project *AEDit Frühe Neuzeit* in cooperation with the *German Text Archive* (*Deutsches Textarchiv,* DTA). They were prepared for machine-readability in accordance with the DTA transcription guidelines. For the "AEDit Frühe Neuzeit" sub-corpus see: Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Edited by the Berlin-Brandenburgische Akademie der Wissenschaften, Berlin 2024, URL: www.deutschestextarchiv.de/search/metadata?corpus=aedit. Accessed: 19 July 2024. For comparability, these models were also generated using the MALLET wrapper from *Gensim.*

2. Methodological Basis

Topic Modeling is a method of Text Mining that aims to access and understand the content of extensive text corpora (for an introductory overview: Blei 2012 a; b; Brett 2012). Unlike classification algorithms where categories are explicitly specified *(supervised machine learning),* Topic Modeling is based on a generative probabilistic modeling process *(unsupervised machine learning).* In this process, the *categories* or *topics* are derived directly from the data. The method resembles traditional indexing practices that have been used since the 18th century to efficiently access certain text units; however, it differs in its approach: instead of fixed keywords, heterogeneous word clusters are generated through probabilistic calculations (Piper 2018, 66–75; see Fig. 1 for an example).

According to *Latent Dirichlet Allocation* (LDA; Blei et al. 2003), the classic Topic Modeling process assumes that the documents of an extensive corpus are made up of different proportions of a fixed set of themes. Furthermore, these themes can be reconstructed as latent, i.e., hidden, linguistic structures or patterns from the text data via the generation of topics (see Blei 2012b, 78–82 for details on the assumptions and the modeling process). To illustrate this, let us assume that we have access to a digitally available library of theological works on Christianity, the content categorization of which via keywords has been lost. The works could contain references to the Trinity, salvation, ethics and morality, as well as biblical exegesis. Topic Modeling enables a reconstruction of these latent content categories. However, Topic Modeling does not generate specific keywords, such as “Trinity,” but groups of words (e.g., “God, Jesus, Spirit, Father, Son, Holy, Trinity …”) that occur together statistically often in the individual documents. The aim is therefore to identify groups of words which, by interpreting their composition, provide an overview of the content structure of the library and its individual works.

Fig. 1 is an exemplary selection of topics in the history of religion for book reviews published on HSoz-Kult; visualization form: word clouds with a weighting of the words according to the relevance for the topic.

In the first step, each word in the works is randomly assigned to a topic. Likewise, a random combination of topics is assigned to each work. In the next step, these initial assignments are checked. The frequency and co-occurrence of a word with other words are used to evaluate whether the current topic assignment is appropriate or whether the word aligns more with a different word cluster. The same method applies to the individual documents: a work in which the words “Jesus”, “holy”, “grace”, “forgiveness”, “sin” and “redemption” occur frequently could, for example, be about the concept of salvation, but was perhaps initially assigned to the topic of “Trinity”. These false assignments are then updated.[2] This process is repeated many, often thousands of times (so-called iterations), until the corpus is structured in a “meaningful” way. In this case, meaningful means that hardly any assignment changes are necessary because the model has stabilized.[3]

Finally, the method produces a statistical model of the library that enables the assignment of individual works to theological topics and ensures efficient orientation within the corpus. This model is represented by two forms of *outputs.* On the one hand, the sources are usually represented as a *document-topic matrix*, i. e., a table documenting the topic weightings for each document. On the other hand, a *topic-word matrix* is generated analogously, breaking down the percentage weighting of the individual words for the individual topics (see also Althage 2022, 260 f.). In so doing, one can abstract from the specific works and extract certain relevant features of the individual texts (i. e., the statistically relevant patterns in language use) as numerical representations, with the aim of understanding the content of the corpus.

Quantitative text analysis methods such as Topic Modeling, which process texts in the form of numerical representations, may initially seem unfamiliar to fields typically engaged in qualitative text-hermeneutic research, such as theology. However, with their macro-analytical approach (cf. Jockers 2013; Graham et al. 2016) these quantitative forms of analyses offer new perspectives on objects of research. There are numerous conceivable applications for theology. Above all, this method is fruitful for analyzing dominant themes, discourses, or concepts – for instance, in sermons, letters, works of the Church Fathers, or scholarly literature. These methods allow one to examine how focal points change over time as well as how different themes or concepts relate to one another. By analyzing texts from different religious groups or authors, one could identify differences and similarities in theological perspectives; various types of texts could also be examined in terms of their linguistic and thematic characteristics.

2 This “assignment” of a topic to a document is expressed as a probability value, which says something about the likelihood of this word cluster occurring in the document or overall corpus.

3 With most Topic Modeling algorithms, one must determine how many topics to be generated for a corpus and in how many repetitions (iterations); one is advised to try different configurations depending on the corpus and the expected diversity of topics (see also section 4).

These possible applications (see also Althage 2022, 259f.) arise from Topic Modeling's ability to process texts as data and thus perform a systematic and scalable analysis. In traditional research contexts, samples or case studies are typically used for an exemplary investigation. Conversely, computational methods can be applied to arbitrarily large source corpora with sufficient computing capacity, thereby also extending the periods of investigation. Given the human cognitive process, extensive corpora are difficult to analyze with consistent examination and relevance criteria, thus what is extracted from the sources develops dynamically and is influenced by a variety of factors (keyword: hermeneutic circle). On the other hand, the computer easily processes very large amounts of data systematically and consistently. The generated topics are provided solely from the data and are not based on categories previously defined based on certain presuppositions.[4] An additional advantage is that Topic Modeling can be applied to any language and therefore to any source data. Through a systematic approach, Topic Modeling enables an in-depth analysis of not just one, but thousands of documents, allowing for the identification and interpretation of hidden thematic structures, thereby gaining a deeper understanding of the particularities of the research object and challenging preconceived assumptions.

3. Topics: Definition and Epistemological Limits

In view of the previously outlined scope of application of Topic Modeling, the term *topic* needs to be defined more precisely. A clearly defined term will help to prevent any misconceptions about the knowledge potential of this method. As with many Text Mining methods, Topic Modeling is mainly based on counting word frequencies. In this context, a topic is a probability distribution across the vocabulary of the text collection that describes the co-occurrence of certain words (Blei 2012b, 78). Although the term "topic" may suggest a resemblance to "Topik" or "Topoi" (Piper 2018, 66–75; Horstmann 2018, 4–7), the term does not carry any epistemological implications beyond the probabilities of co-occurrence, meaning the joint appearance of words (cf. Blei et al. 2003, 996, note 1; Althage 2022, 267; see also Shadrova 2021). Within the context of humanities research, however, applying Topic Modeling is associated with two assumptions: First, that of topic coherence, which states that the terms assigned to a topic should have a thematic or conceptual relationship; and second, the assumption of stability of meaning, according to which a given topic, if assigned to different documents, should have the same meaning or relevance for all these documents (Schmidt 2012, 49).

4 At the same time, however, this also means that the type and scope of data preprocessing has a significant influence on the modeling result. Cf. section 4.2.

However, Topic Models do not *understand* the meaning and concepts that people associate with the words of a text, given the computer is “semantically blind” in this respect (Schwandt 2018, 108. 133). Accordingly, Benjamin Schmidt critically pointed out that topics are not inherently meaningful but become so through our interpretation (Schmidt 2012; see also Horstmann 2018, 10). David Blei noted that topics could resemble themes because words that often co-occur tend to be part of the same thematic field (Blei 2012a, 9). As such, claiming these topics as themes or discourses is based on the principle of distributional semantics (Piper 2018, 13; Schöch 2017, 14). Distributional semantics formalizes the assumption that the meaning of words arises from their co-occurrence frequency with other words in a particular context. The context can be a document, a paragraph, or even a single sentence. To interpret text data on a “semantic level”, these frequency relationships between words are numerically represented by, for example, coordinates in a vector space, making them computationally processable (Turney & Pantel 2010; Blei 2012a, 9; Piper 2018, 13–18; see also Althage 2022, 266 f.).

Despite topics being frequently equated with *themes* or other semantic categories, they are not synonymous (Uglanova & Gius 2020, 72). This is also reflected in the fact that a topic model usually also includes word clusters – sometimes heavily weighted – that describe more general stylistic properties of a specific text type (*meta-topics,* see Fig. 2) or indicate heterogeneity in the text data, which is reflected, for example, in language-specific topics (cf. Fig. 3; on various topic forms, see Boyd-Graber et al. 2014, 234–237; Schöch 2017, 23–26; Althage 2022, 267–269). The latter, referred to as *Noisy Topics,* could serve as a starting point for further preprocessing of the text data (see section 4.2). In the context of the aforementioned assumptions, it should also be noted that a topic with the same weighting in two different documents can also have completely different emphases at the word level.[5] Topics are thus not independent epistemological units with a fixed semantic core, but rather a hermeneutic instrument (Rockwell & Sinclair 2016). Topics enable a structured approach to large amounts of text yet always require interpretation performed within the context of the underlying sources, whereby the assumptions of coherence and stability of meaning in particular should be examined.

5 See, e.g., the reviews at https://www.hsozkult.de/publicationreview/id/reb-25382 and https://www.hsozkult.de/publicationreview/id/reb-26856, which, with a distribution of 32 % each for topic 42 (see Fig. 1), reflect different conceptual dimensions of the topic. Both addresses were accessed on 18 June 2024.

Fig. 2 Examples of *Metatopics* from the book reviews of H-Soz-Kult

Fig. 3 Examples of *Noisy Topics* from the book reviews of H-Soz-Kult

4. Topic Modeling Workflow

The concrete application of Topic Modeling in a research context requires a carefully constructed, critically reflected, and documented workflow (see Fig. 4). The workflow includes the selection of a suitable method, the preparation of the text data *(preprocessing)* for the generation of Topic Models and the evaluation of the results, considering various configurations of preprocessing and Topic Modeling. Generally, this process is iterative where it is possible to go back and forth between the individual processing steps to optimize the modeling results regarding the research question. Once an appropriate Topic Model has been found, there are various visualization options for the results, from simple word lists and word clouds to bar graphs, line charts, or scatter plots to illustrate characteristics and developments, and heat maps for correlations or networks for relationships between the clusters. This section focuses on the selection of a suitable algorithm, the preprocessing of the data, and the evaluation of the results.

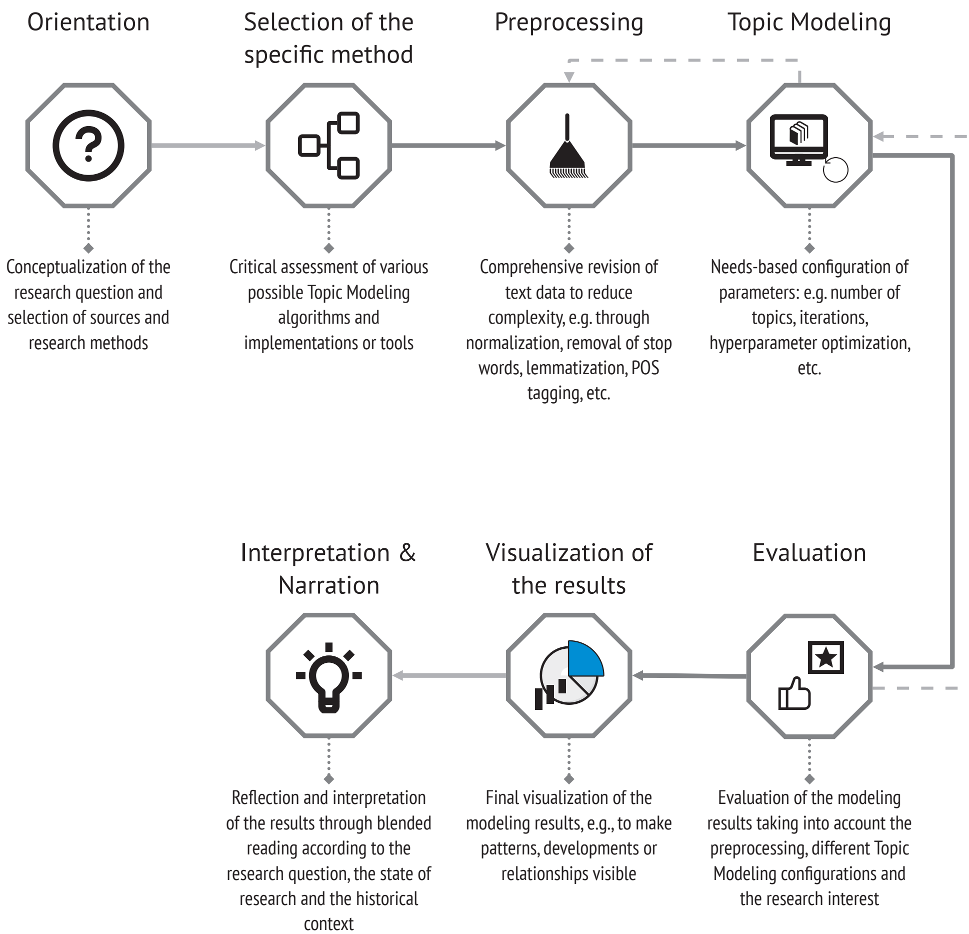


Fig. 4 Schematic Topic Modeling workflow

4.1 Selection of the Topic Modeling Method

At the beginning of the research project, it should be determined which algorithm in which implementation is suitable for a given research question (cf. Jelodar et al. 2019; Vayansky & Kumar 2020; Churchill & Singh 2022). This decision should be based on a comparison of various approaches and their respective results (Fig. 5 is an exemplary aid for decision-making). Various factors must be considered, e.g., the consistency of the theoretical-methodological assumptions of the procedure with one's own research objectives and the available configuration options (from the number of topics to hyperparameter optimization) and their effects on the output. The aim is to critically examine the potentials and limitations of the methods under consideration.

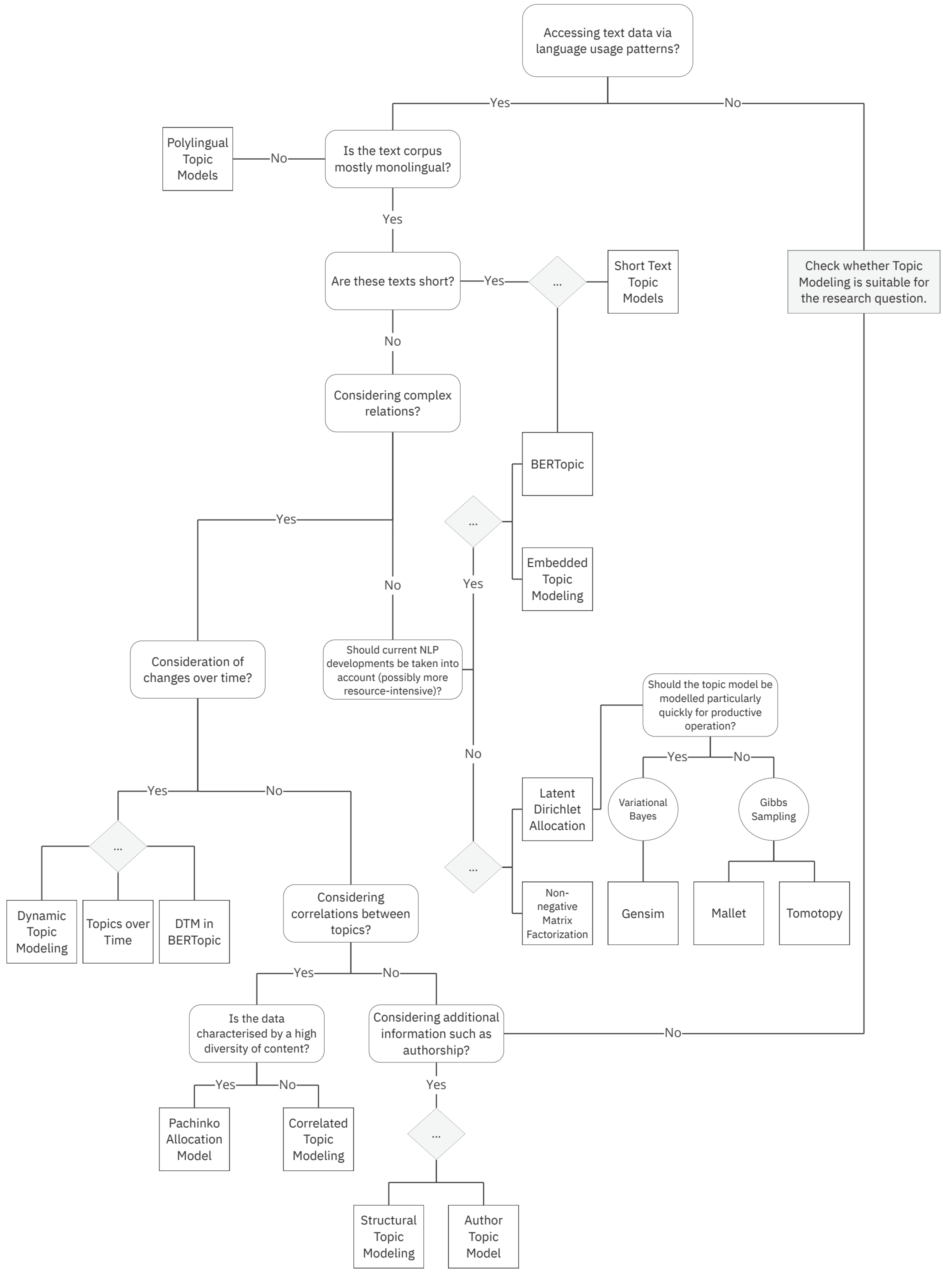


Fig. 5 Exemplary decision tree (building on Vayansky & Kumar 2020, esp. 14, Fig. 8; Churchill & Singh 2022; Jelodar et al. 2019); some key questions about the goal or the characteristics of the research object can help to select the appropriate method or tool.*

* On *Non-negative Matrix Factorization* see: Lee & Seung 1999; *Topics over Time:* Wang & McCallum 2006; *Pachinko Allocation Model:* Li & McCallum 2006; on *Embedded Topic Modeling* for example: Dieng et al. 2020.

Among the numerous options, LDA (Blei et al. 2003) has proven to be particularly popular in the Digital Humanities and is also implemented in numerous ready-to-use tools and programming libraries.[6] This method has been used successfully as a heuristic tool in a variety of disciplines to explore extensive text collections. In the Digital Humanities, the method is also used for historical studies that extend over longer periods of time (such as, Wehrheim et al. 2022; Snickars 2022; Grant et al. 2021). However, since LDA does not consider the temporal and relational dimension of the data or its contextuality in the modeling process, the information must be applied to the model retrospectively (Althage 2022). In contrast, methods like *Dynamic Topic Modeling* (DTM; Blei & Lafferty 2006; Grootendorst 2022; for an application of the method, cf. Guldi 2019) take the temporality of the topics into account during the modeling process, allowing them to show how, for example, discourses emerge, develop, and disappear over time.

Should the focus be on the relationships between different clusters rather than the temporal dimensions, there are suitable methods like *Correlated Topic Modeling* (CTM; Lafferty & Blei 2005; Blei & Lafferty 2007). CTM can show how different topics correlate with each other. *Structural Topic Modeling* (STM; Roberts et al. 2014; Küsters & Garrido 2020), on the other hand, enables one to model topics in relation to specific contextual information, which is particularly useful when investigating the influence of factors such as gender, social group affiliation, or genre on topic formation. The influence of authorship on the topic model can also be investigated using STM, but special *Author Topic Models* have also been developed (Rosen-Zvi et al. 2004).

The versatility of Topic Modeling is evident in its applicability to different types of text, from scientific articles to historical documents and social media posts. However, for shorter texts, for instance tweets or titles of works, specialized models such as *Short Text Topic Models* (Cheng et al. 2014; Zuo et al. 2016; Zhao et al. 2021) may be more suitable. For multilingual text collections, *Polylingual Topic Models* (Mimno et al. 2009) or *BERTopic* (Grootendorst 2022) can be used to identify thematic consistencies across different languages.

After providing an overview of various Topic Modeling methods and their application, the question arises as to how these models practically can be integrated into the research process. There are various options: for example, ready-to-use tools such as the *TopicsExplorer*[7] or *Topics* in *Voyant Tools*[8] can be used. Because the configu-

6 However, it should be noted that there are different implementations of LDA, which can generate different modeling results. For example, the implementations in *MALLET* and *Gensim* differ in terms of their inference algorithms for deriving the topics. The *Gensim* implementation is designed to handle very large amounts of data and focuses on performance; the results can therefore be less coherent. In contrast, *MALLET* requires more computational time to model the topics, but generally produces more coherent and robust modeling results – even with smaller text corpora. Cf. Althage 2022, 261–263; Hodel et al. 2022; Boyd-Graber et al. 2014, 231–233.

7 See https://dariah-de.github.io/TopicsExplorer (Accessed: 19 July 2024).

8 See http://voyant-tools.org (Accessed: 18 June 2024), cf. Rockwell & Sinclair 2016..

ration options have a substantial impact on the results, especially with this method, these software solutions should primarily be seen as an introductory aid to familiarize oneself with the modeling process. While the number of topics to be generated and their iterations can often be freely selected using these tools, more complex components such as hyperparameters, which influence the distribution profile of the topics (Wallach et al. 2009; Boyd-Graber et al. 2014, 233), are hidden in the *Black Box.* The possibilities for evaluating the modeling results or exporting them as reusable data are also limited. Furthermore, it should be noted that Topic Modeling, as we have seen, represents a whole range of algorithms that all pursue, in one way or another, the goal of grouping texts based on their patterns of language use in order to explore their thematic structure.

It is therefore advisable to use more complex solutions such as the widely used framework *MALLET,*[9] the *interactive Leipzig Corpus Miner* (iLCM),[10] or implementations of various algorithms in programming languages such as *Python* or *R,* enabling one to configure the procedures according to one's needs. Libraries in Python such as *Gensim,*[11] *Scikit-Learn,*[12] *Tomotopy,*[13] *BERTopic,*[14] or *OCTIS,*[15] for example, offer several solutions within a single environment.[16] Choosing the appropriate algorithm and implementation is only the first step in a complex process; as we will discuss in the next chapter, the careful preparation of the text corpus is essential for historical and stylistically diverse text data.

4.2 Preprocessing

Topic Modeling can, in principle, be applied to any text from any language. Although especially for historical research disciplines, one should consider that the methods presented above were usually developed and tested using text data corresponding to modern languages, which are more standardized than medieval or early modern texts. Additionally, literary texts with their numerous stylistic peculiarities can also pose a challenge in this context (see on this for example Uglanova & Gius 2020). LDA, e. g., was tested using, among others, English language news and scientific articles (Blei et al. 2003). As the method makes regularities visible, a corpus linguisti-

9 See http://mallet.cs.umass.edu (Accessed: 18 June 2024), cf. McCallum 2002.
10 See https://ilcm.informatik.uni-leipzig.de (Accessed: 18 June 2024), cf. Niekler et al. 2023.
11 See https://radimrehurek.com/gensim (Accessed: 18 June 2024), cf. Řehůřek & Sojka 2010.
12 See https://scikit-learn.org/stable/index.html (Accessed: 18 June 2024).
13 See https://bab2min.github.io/tomotopy (Accessed: 18 June 2024).
14 See https://maartengr.github.io/BERTopic/index.html (Accessed: 19 July 2024), cf. Grootendorst 2022.
15 See https://github.com/MIND-Lab/OCTIS (Accessed: 18 June 2024), cf. Terragni et al. 2021.
16 A look at the documentation of the libraries provides information about the configuration options and initial sample code. Usually, there are also numerous useful tutorials available online.

cally and orthographically homogeneous is more reliable to model than 17th century funeral writings, which were not yet subject to comparable written language rules and can have different spellings for the same concepts as well as numerous Latin remarks and quotations. The more complex and varied the sources, the less consistent and predictable the results of the modeling may potentially be.

Tab. 1 Exemplary comparison of a selection of topics before and after initial preprocessing (17th century funeral sermons, "AEDit Frühe Neuzeit")

Before Preprocessing	After Preprocessing (Tokenization, Removal of Punctuation Marks and Numbers, Lemmatization, POS tagging, Lowercasing)
und der die das zu mit auch er nicht den dem ist sie von ein wie des sich Gott daß	kind eltern job söhnlein kinderlein kindlein lieb gerecht taufe bräutigam töchterlein braut gerechtigkeit item de christus matt arm justitia himmlisch
Frau Kinder Mutter Adeligen Eltern Kind Adelige Gn. Edlen/E. J. Kinderlein liebes Weib Eltern/Rahel geborene Söhnlein Kindlein Job	frau lieb mutter adelig junker weib adelige witwe edl gn geboren rahel herz schmerz trost schwester gestreng kind kreuz augenlust
Prediger Lehrer & Kirchen Amt M. Zuhörer Stadt ad Anno D. Prediger/c. treuen Gemeine Schulen Fürstlichen Fürstl. treue	prediger kirche lehrer amt jahr zuhörer wort treu prophet groß lehre schule stadt apostel anno ehrwürdig predigt knecht fürstl mann
Dann dann wann wider sonder „ Vers deren lang dieselbige Leibs Tods Kapitel gern Edlen Arzt Sara Sohns seliger dieweil	christus arzt jesus arznei kreuz kapitel doktor luc matth apotheke christi wunde joh apotheker volk hiob medikus leiden heiland jude
Sie Er daß die der Ihr als eine Ihm von zu Frau Die Ich dem den GOTT sich Der Ihre	frau seele hoch himmel welt mutter tod haus freude träne auge herrlichkeit ps leben liebe braut ehre vater land tugend

To filter out the content characteristics of the texts (see Tab. 1), it is advisable to reduce the complexity of the text data by standardizing and normalizing the vocabulary; this work step is called *preprocessing* (cf. Maier et al. 2018, 97f. 100–102. 110). The type, scope, and sequence of the individual processing steps are significant and depend on the chosen method, as well as on the type of source to be processed and the respective research question. When selecting and arranging the individual processing steps, one has to consider the language and the degree of standardization and normalization. While the available resources for modern languages are increasing, the availability for historical languages is still limited, which can result in more complex preprocessing. Be it modern or historical texts, one must proceed carefully when preparing the data and documenting the individual decisions in this often iterative process to ensure the traceability of the procedure.

Tokenization is one of the mandatory preparatory steps of numerous text analysis methods. This involves breaking down the text into smaller units, also known as *tokens,* which can then be processed, counted, compared, and recombined. Tokenization is usually performed on words. While humans can intuitively recognize lexical units, this process must be explicitly formalized for the computer. Depending on the language, this computation poses its own challenges. The handling of multi-word units is particularly relevant, which are sometimes but not always identified by hyphens, as in the case of "Holy Spirit." Typically, the connection between the two words would be removed during tokenization ("Holy," "Spirit"), so that the individual word components are processed independently of each other. Therefore, it is essential to reflect on what should be considered as the unit of investigation in the respective research context. For such *Natural Language Processing* tasks, numerous established tools already exist.[17] Moreover, the modeling of *bigrams* (word pairs) and *trigrams* (word triples) can also assist in reassembling phrases composed of co-occurring terms into a single token, thereby partially preserving the local reference.

There is no prescribed path for further preprocessing; however, several steps have been established that can be used modularly and tailored to the data corpus, depending on the intended purpose (see Fig. 6). One such step is *segmentation.* Particularly when dealing with corpora made up of very extensive individual documents (e.g., a corpus of books), the documents can be broken down into smaller units (e.g., at chapter or paragraph level). The removal of punctuation and the so-called stop words has also become common practice (see e.g. Schofield et al. 2017 for a critical perspective on the removal of stop words). The stop words are function words like "the," "in," "at," etc., which, as they occur very frequently in texts, would dominate the topics as statistically very prominent characteristics of texts (see Tab. 1). Even if these words are of central grammatical importance, they do not necessarily belong to the words that carry meaning and could make working with the Topic Model more difficult. Depending on the language, there are various stop word lists that can be reused through programming libraries, such as NLTK, and extended manually with additional terms specific to the corpus. The contents of these lists should be checked to ensure that relevant words to one's project are not removed.

Methods like *TF-IDF* or *Part-of-Speech Tagging* (POS tagging) offer a more systematic approach to selecting relevant and meaningful words than stop word lists. With TF-IDF, one can identify tokens that are characteristic of a particular document or group of documents and, in contrast, give less weight to words that occur particularly frequently in many documents (such as function words, but also other corpus-specific terms) and filter them out accordingly (Klinke 2017, 274 f.). POS tagging, on the other hand, automatically determines the word types of the lexical units. In this way, spe-

17 In *Python,* e.g., see NLTK. Natural Language Toolkit, URL: https://www.nltk.org; spaCy. Industrial-Strength Natural Language Processing, URL: https://spacy.io. Both addresses were accessed on 18 June 2024.

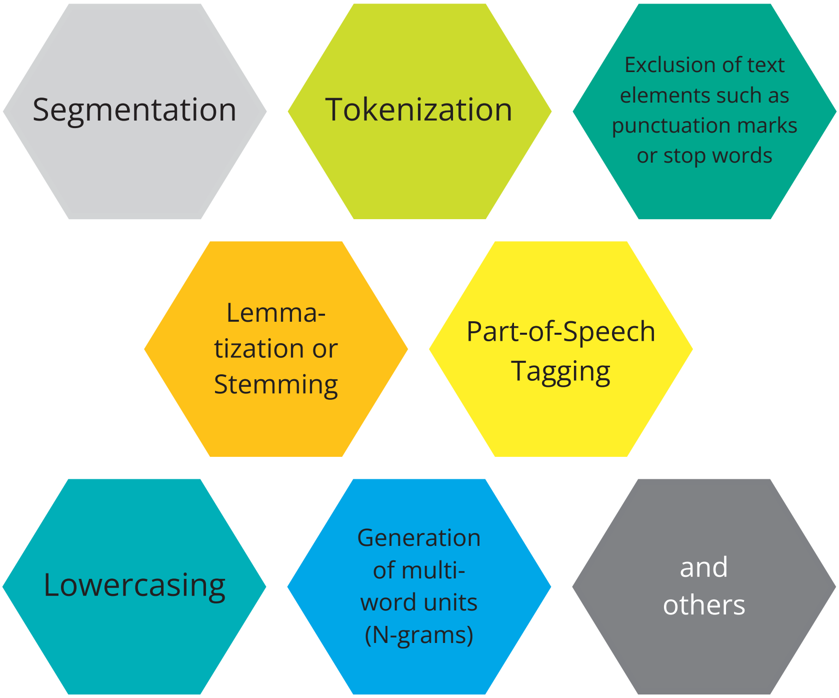


Fig. 6 The preparation of text data can be composed of various modular steps depending on the characteristics and the quality of the data as well as the research objectives.

cific word types can be selected for analysis that are assumed to have a meaningful function in texts (e.g., nouns, verbs, or adjectives; see for example, Schöch 2017, 17).

To minimize the variations in word forms and enable the modeling of more coherent topics, as well as to facilitate data processing, in addition to *lowercasing* (i.e., the lowercasing of all tokens), *lemmatization* has proven to be efficient. This approach reduces inflected single word forms to their base form (holier → holy, went → go). *Stemming* is also not uncommon, especially in English contexts. The individual words are reduced to their stem or root by truncating the word endings (e.g., Protestant → Protest, Protestantism → Protestant[18]). Stemming results in tokens that do not necessarily reflect a valid lexical entry in a language and can therefore be much more difficult to interpret (Schofield & Mimno 2016). If the sole aim is to index a corpus efficiently, then the latter approach might be appropriate; however, lemmatization is the preferred option for research projects aimed at interpretation.

18 Editor's note: A different example was chosen in the original text, which does not work so well in English: Christian, Christ → Christ, Christianity → Christian.

Each step of processing has a direct effect on the respective modeling result and thus on what is intended to be interpreted.[19] The procedure should therefore not only be documented, but also integrated into the evaluation of the Topic Models, taking into account the research question and knowledge objectives.

4.3 Evaluation

Evaluating the Topic Models is essential to ensure the quality and relevance of the generated clusters. This is especially advisable as there is a risk of succumbing to *confirmation bias* with such methods, i.e. processing the modeling and the resulting data until a desired or expected outcome is achieved (Shadrova 2021, 5, 16 f.). Although, as Maria Antoniak pointed out, Topic Modeling is not about generating the one "correct" perspective of the text corpus, but about supporting a qualitative investigation by discovering one of many possible "interpretative lenses" through which sources can be understood (Antoniak 2022), it is advisable to also include mathematical evaluation metrics in addition to the qualitative examination of the topics in terms of their interpretability and representativeness (a good introduction is provided by Boyd-Graber et al. 2014, 233 f. 237–243; Churchill & Singh 2022, 5–9).

For qualitative evaluation, *Blended Reading* can be applied as an evaluation mode (Stulpe & Lemke 2016). This approach combines the results of the machine learning process *(Distant Reading)* with human reading and interpretation *(Close Reading).* By reading representative documents or text passages for each topic, one can evaluate their interpretability and representativeness. Or by comparing the most important words and phrases assigned to the topics, the granularity of the model can be determined.[20] It can also be useful to check whether there is a Ground Truth for the text corpus or whether it is feasible to create one. This means, for example, an already existing manual classification that the model can be compared against.[21] In this case, metrics such as *accuracy, precision, recall,* and *F-score* can be used to evaluate the performance of the model (Churchill & Singh 2022, 5–9; Klinke 2017, 269 f.).

Since there is usually no such Ground Truth when using Topic Modeling, a number of other metrics have been established that can be used to evaluate the modeling results (Churchill & Singh 2022, 6–8; Boyd-Graber et al. 2014, 233 f. 237–243), including:

19 Newer methods such as *BERTopic*, promise to be able to dispense with preprocessing by using the latest language models. However, it remains to be seen how well this works for historical languages.

20 The number of topics has an impact on the granularity of the model. Too high a number potentially leads to overlapping, redundant clusters, while too low a number results in clusters that are too heterogeneous, see also Schöch 2017, 20, note 7.

21 In case of H-Soz-Kult, for example, there is a manual classification according to themes, regions, and epochs, which provides a good orientation for the evaluation of the modeling results.

- Coherence measures can be used, for example, to measure how well the (top) words assigned to the topics fit together: the higher the coherence value, the more semantically coherent the topics and thus interpretable the word clusters are in theory. Tools such as *Gensim* in *Python* offer functions for calculating coherence, which can also be used to determine which number of topics is appropriate for a given research topic.[22]
- Perplexity, on the other hand, can be used to assess how well the topic model can predict new, unseen documents. A lower perplexity value is typically better, but this value alone is often not sufficient to assess the quality of a model.
- The exclusivity or uniqueness of the (top) words assigned to the topics can also be measured for the respective topics in order to assess the distinctiveness of the word clusters.

The above are just a few of the available options for evaluating topic models. Since these evaluation metrics do not always correlate positively with human assessments of the modeling results (e.g. Hoyle et al. 2021; Uglanova & Gius 2020), they should always be used in addition to the qualitative manual interpretation by the researchers with their domain knowledge and with due consideration of the specific research question and the characteristics of the text corpus.

5. Concluding Remarks

In conclusion, it can be stated that Topic Modeling can be a valuable addition to the methodological landscape of theology. As a “statistical lens” that formalizes the knowledge, theories, and assumptions of theology (after Blei 2012a, 8), it can provide new data-driven perspectives on sources and research debates. Although Topic Models do not, by themselves, provide conclusions to specific research questions, they can be used to explore hypotheses and test them against the individual sources. Topic Modeling does not replace textual hermeneutic approaches, but rather expands the study of sources with an additional set of tools. Thus, it bridges the gap between traditional hermeneutic approaches and modern, data-based methods. It enriches the analytical repertoire of the humanities and creates new avenues for the systematic and critical examination of extensive text corpora. It is to be hoped that this approach will serve as a starting point for further investigations and discussions within theology.

22 With *Hierarchical Dirichlet Process* (HDP), an extension of LDA, a method was developed making it possible to derive the number of topics from the corpus data (Teh et al. 2006).

References

Althage, M. (2022). Potenziale und Grenzen der Topic-Modellierung mit Latent Dirichlet Allocation für die Digital History. In K.D. Döring, S. Haas, M. König & J. Wettlaufer (Eds.), *Digital History. Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft* (pp. 255–277). Berlin/Boston: De Gruyter Oldenbourg [= *Studies in Digital History and Hermeneutics,* 6]. DOI: https://doi.org/10.1515/9783110757101-014 (Accessed: 18 June 2024).

Ead. (2023). Digitale Methoden kritisch reflektieren. Die Erweiterung des Werkzeugkastens der Historiker:innen. In P. Trilcke, A. Busch & P. Helling (Eds.), *DHd 2023. Open Humanities Open Culture. 9. Tagung des Verbands Digital Humanities im deutschsprachigen Raum.* Trier/Luxemburg: Zenodo. DOI: https://doi.org/10.5281/zenodo.7711522 (Accessed: 18 June 2024).

Antoniak, M. (2022). Topic Modeling for the People. In *Blog by M. Antoniak.* URL: https://maria-antoniak.github.io//2022/07/27/topic-modeling-for-the-people.html (Accessed: 18 June 2024).

Blei, D.M. (2012a). Topic Modeling and Digital Humanities, *Journal of Digital Humanities,* 2(1), 8–11. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei (Accessed: 18 June 2024).

Id. (2012b). Probabilistic topic models, *Communications of the ACM,* 55(4), 77–84. DOI: https://doi.org/10.1145/2133806.2133826 (Accessed: 18 June 2024).

Id., & Lafferty, J.D. (2006). Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning – ICML '06* (pp. 113–120). New York: ACM Press. DOI: https://doi.org/10.1145/1143844.1143859 (Accessed: 18 June 2024).

Eid. (2007). A correlated topic model of Science, *The Annals of Applied Statistics* 1(1), 17–35. DOI: https://doi.org/10.1214/07-AOAS114 (Accessed: 18 June 2024).

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research,* 3, 993–1022. DOI: https://dl.acm.org/doi/10.5555/944919.944937 (Accessed: 21 July 2024).

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and Feeding of Topic Models. Problems, Diagnostics, and Improvements. In E.M. Airoldi, D.M. Blei, E.A. Erosheva & S.E. Fienberg (Eds.), *Handbook of Mixed Membership Models and Their Applications* (pp. 225–254). London/New York: Chapman and Hall/CRC. DOI: https://doi.org/10.1201/b17520-21 (Accessed: 18 June 2024).

Brett, M.R. (2012). Topic Modeling. A Basic Introduction, *Journal of Digital Humanities,* 2(1), 12–16. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett (Accessed: 18 June 2024).

Bunout, E., & von Lange, M. (2019). Nibbling at Text. Identifying Discourses on Europe in a Large Collection of Historical Newspapers Using Topic Modelling. In *C2DH | Luxembourg Centre for Contemporary and Digital History.* URL: https://www.c2dh.uni.lu/thinkering/nibbling-text-identifying-discourses-europe-large-collection-historical-newspapers-using (Accessed: 18 June 2024).

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM. Topic Modeling over Short Texts, *IEEE Transactions on Knowledge and Data Engineering,* 26(12), 2928–2941. DOI: https://doi.org/10.1109/TKDE.2014.2313872 (Accessed: 18 June 2024).
Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling, *ACM Computing Surveys,* 54(10s), 1–35. DOI: https://doi.org/10.1145/3507900 (Accessed: 18 June 2024).
Crane, G. (2006). What Do You Do with a Million Books?, *D-Lib Magazine,* 12(3), no pag. DOI: https://doi.org/10.1045/march2006-crane (Accessed: 18 June 2024).
Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic Modeling in Embedding Spaces, *Transactions of the Association for Computational Linguistics,* 8, 439–453. DOI: https://doi.org/10.1162/tacl_a_00325 (Accessed: 21 July 2024).
Graham, Sh., Milligan, I., & Weingart, S. (2016). *Exploring Big Historical Data. The Historian's Macroscope.* London: Imperial College Press.
Grant, Ph., Sebastian, R., Allassonnière-Tang, M., & Cosemans, S. (2021). Topic Modelling on Archive Documents from the 1970s. Global Policies on Refugees, *Digital Scholarship in the Humanities,* 36(4), 886–904. DOI: https://doi.org/10.1093/llc/fqab018 (Accessed: 18 June 2024).
Graves, M. (2022). Computational Topic Models for Theological Investigations, *Theology and Science,* 20(1), 69–84. DOI: https://doi.org/10.1080/14746700.2021.2012922 (Accessed: 18 June 2024).
Grootendorst, M. (2022). BERTopic. Neural topic modeling with a class-based TF-IDF procedure. Online: *arXiv.* DOI: https://doi.org/10.48550/arXiv.2203.05794 (Accessed: 18 June 2024).
Guldi, J. (2019). Parliament's Debates about Infrastructure. An Exercise in Using Dynamic Topic Models to Synthesize Historical Change, *Technology and Culture,* 60(1), 1–33. DOI: https://doi.org/10.1353/tech.2019.0000 (Accessed: 18 June 2024).
Hodel, T., Möbus, D., & Serif, I. (2022). Von Inferenzen und Differenzen. Ein Vergleich von Topic-Modeling-Engines auf Grundlage historischer Korpora. In S. Gerlek, S. Kissler, Th. Mämecke, & D. Möbus (Eds.), *Von Menschen und Maschinen. Mensch-Maschine-Interaktionen in digitalen Kulturen* (pp. 185–209). Hagen: Hagen University Press. DOI: https://doi.org/10.57813/20220623-153139-0 (Accessed: 21 July 2024)
Horstmann, J. (2018). Topic Modeling, *ForText. Literatur Digital Erforschen,* 1–16. URL: https://fortext.net/routinen/methoden/topic-modeling (Accessed: 18 June 2024).
Hoyle, A., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J., & Resnik, Ph. (2021). Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. In *Advances in Neural Information Processing Systems,* 34 (pp. 2018–2033). URL: https://proceedings.neurips.cc/paper_files/paper/2021/hash/0f83556a305d789b1d71815e8ea4f4b0-Abstract.html (Accessed: 18 June 2024).
Jelodar, H., Wang, Y., Yuan, Ch., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic Modeling. Models, Applications, a

Survey, *Multimedia Tools and Applications,* 78(11), 15169–15211. DOI: https://doi.org/10.1007/s11042-018-6894-4 (Accessed: 18 June 2024).

Jockers, M.L. (2013). *Macroanalysis. Digital Methods and Literary History.* Urbana/Chicago/Springfield: University of Illinois Press.

Klinke, H. (2017). Information Retrieval. In F. Jannidis, H. Kohle, & M. Rehbein (Eds.), *Digital Humanities. Eine Einführung* (pp. 268–278). Stuttgart: J.B. Metzler. DOI: https://doi.org/10.1007/978-3-476-05446-3_19 (Accessed: 18 June 2024).

Küsters, A., & Garrido, E. (2020). Mining PIGS. A Structural Topic Model Analysis of Southern Europe Based on the German Newspaper Die Zeit (1946–2009), *Journal of Contemporary European Studies,* 28(4), 477–493. DOI: https://doi.org/10.1080/14782804.2020.1784112 (Accessed: 18 June 2024).

Lafferty, J.D., & Blei, D.M. (2005). Correlated Topic Models. In *Advances in Neural Information Processing Systems,* 18 (pp. 147–154). URL: https://proceedings.neurips.cc/paper_files/paper/2005/hash/9e82757e9a1c12cb710ad680db11f6f1-Abstract.html (Accessed: 18 June 2024).

Lee, D.D., & Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization, *Nature,* 401, 788–791. DOI: https://doi.org/10.1038/44565 (Accessed: 21 July 2024).

Li, W., & McCallum, A. (2006). Pachinko Allocation. DAG-Structured Mixture Models of Topic Correlations. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 577–584). New York: Association for Computing Machinery. DOI: https://doi.org/10.1145/1143844.1143917 (Accessed: 21 July 2024).

Luhmann, J., & Burghardt, M. (2021). Digital Humanities. A Discipline in Its Own Right? An Analysis of the Role and Position of Digital Humanities in the Academic Landscape, *Journal of the Association for Information Science and Technology,* 73(2), 148–171. DOI: https://doi.org/10.1002/asi.24533 (Accessed: 18 June 2024).

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research. Toward a Valid and Reliable Methodology, *Communication Methods and Measures,* 12(2–3), 93–118. DOI: https://doi.org/10.1080/19312458.2018.1430754 (Accessed: 18 June 2024).

McCallum, A.K. (2002). MALLET. A Machine Learning for Language Toolkit. URL: http://mallet.cs.umass.edu (Accessed: 18 June 2024).

Mimno, D. (2012). Computational Historiography. Data Mining in a Century of Classics Journals, *Journal on Computing and Cultural Heritage,* 5(1), 1–19. DOI: https://doi.org/10.1145/2160165.2160168 (Accessed: 18 June 2024).

Id., Wallach, H.M., Naradowsky, J., Smith, D.A., & McCallum, A. (2009). Polylingual Topic Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 880–889). Singapore: Association for Computational Linguistics. URL: https://aclanthology.org/D09-1092 (Accessed: 21 July 2024)

Niekler, A., Kahmann, Ch., Burghardt, M., & Heyer, G. (2023). The Interactive Leipzig Corpus Miner. An Extensible and Adaptable Text Analysis Tool for Content Analysis. *Publizistik,* 68, 325–354. DOI: https://doi.org/10.1007/s11616-023-00809-4 (Accessed: 18 June 2024).
Nunn, Ch. A. (2022). Das Thema patristischer Ethik. Versuch einer Annäherung durch Distanz am Beispiel der Briefe des Augustinus von Hippo, *Journal of Ethics in Antiquity and Christianity,* 4, 31–51. DOI: https://doi.org/10.25784/jeac.v4i0.1011 (Accessed: 18 June 2024).
Piper, A. (2018). *Enumerations. Data and Literary Study.* Chicago/London: The University of Chicago Press.
Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 46–50). Valletta, Malta: ELRA.
Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses, *American Journal of Political Science,* 58(4), 1064–1082. DOI: https://doi.org/10.1111/ajps.12103 (Accessed: 18 June 2024).
Rockwell, G., & Sinclair, S. (2016). *Hermeneutica. Computer-Assisted Interpretation in the Humanities.* Cambridge, MA/London: The MIT Press.
Rosen-Zvi, M., Griffiths, Th., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for Authors and Documents. In *UAI '04. Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (pp. 487–494). Arlington: AUAI Press. DOI: https://dl.acm.org/doi/10.5555/1036843.1036902 (Accessed: 21 July 2024).
Schmidt, B. M. (2012). Words Alone. Dismantling Topic Models in the Humanities, *Journal of Digital Humanities,* 2(1), 49–65. URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt (Accessed: 18 June 2024).
Schöch, Ch. (2017). Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama, *Digital Humanities Quarterly,* 11(2), 1–53. URL: https://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html (Accessed: 19 July 2024).
Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling Out the Stops. Rethinking Stopword Removal for Topic Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2. Short Papers* (pp. 432–436). URL: https://www.aclweb.org/anthology/E17-2069 (Accessed: 18 June 2024).
Schofield, A., & Mimno, D. (2016). Comparing Apples to Apple. The Effects of Stemmers on Topic Models, *Transactions of the Association for Computational Linguistics,* 4, 287–300. DOI: https://doi.org/10.1162/tacl_a_00099 (Accessed: 18 June 2024).
Schwandt, S. (2018). Digitale Methoden für die Historische Semantik. Auf den Spuren von Begriffen in digitalen Korpora, *Geschichte und Gesellschaft. Zeitschrift für historische Sozialwissenschaft,* 44(1), 107–134. DOI: https://doi.org/10.13109/gege.2018.44.1.107 (Accessed: 18 June 2024).

Shadrova, A. (2021). Topic Models Do Not Model Topics. Epistemological Remarks and Steps towards Best Practices, *Journal of Data Mining & Digital Humanities,* 1–28. DOI: https://doi.org/10.46298/jdmdh.7595 (Accessed: 18 June 2024).

Simmler, S., Vitt, Th., & Pielström, S. (2019). Topic Modeling with Interactive Visualizations in a GUI Tool. In *Proceedings of the Digital Humanities Conference.* Utrecht. Tool: https://dariah-de.github.io/TopicsExplorer (Accessed: 18 June 2024).

Snickars, P. (2022). Modeling Media History. On Topic Models of Swedish Media Politics 1945–1989, *Media History,* 28(3), 403–424. DOI: https://doi.org/10.1080/13688804.2022.2079484 (Accessed: 18 June 2024).

Stulpe, A., & Lemke, M. (2016). Blended Reading. Theoretische und praktische Dimensionen der Analyse von Text und sozialer Wirklichkeit im Zeitalter der Digitalisierung. In M. Lemke & G. Wiedemann (Eds.), *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse* (pp. 17–61). Wiesbaden: Springer VS. DOI: https://doi.org/10.1007/978-3-658-07224-7_2 (Accessed: 21 July 2024).

Teh, Y.W., Jordan, M.I., Beal, M.J., & Blei, D.M. (2006). Hierarchical Dirichlet Processes, *Journal of the American Statistical Association,* 101(476), 1566–1581. DOI: https://doi.org/10.1198/016214506000000302 (Accessed: 18 June 2024).

Terragni, S., Fersini, E., Galuzzi, B.G., Tropeano, P., & Candelieri, A. (2021). OCTIS. Comparing and Optimizing Topic models is Simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations* (pp. 263–270). DOI: https://doi.org/10.18653/v1/2021.eacl-demos.31 (Accessed: 18 June 2024).

Turney, P.D., & Pantel, P. (2010). From Frequency to Meaning. Vector Space Models of Semantics, *Journal of Artificial Intelligence Research,* 37(1), 141–188. DOI: https://jair.org/index.php/jair/article/view/10640 (Accessed: 18 June 2024).

Uglanova, I., & Gius, E. (2020). The Order of Things. A Study on Topic Modelling of Literary Texts. In *Proceedings of the Workshop on Computational Humanities Research* (CHR 2020) (pp. 57–76). URL: https://ceur-ws.org/Vol-2723/long7.pdf (Accessed: 18 June 2024).

Vayansky, I., & Kumar, S.A.P. (2020). A Review of Topic Modeling Methods, *Information Systems,* 94(101582), 1–15. DOI: https://doi.org/10.1016/j.is.2020.101582 (Accessed: 18 June 2024).

Völkl, Y., Sarić, S., & Scholger, M. (2022). Topic Modeling for the Identification of Gender-Specific Discourse. Virtues and Vices in French and Spanish 18th Century Periodicals, *Journal of Computational Literary Studies,* 1(1), 1–27. DOI: https://doi.org/10.48694/jcls.108 (Accessed: 18 June 2024).

Wallach, H.M., Mimno, D., & McCallum, A. (2009). Rethinking LDA. Why priors matter. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (pp. 1973–1981). New York: Curran Associates Inc. DOI: https://dl.acm.org/doi/10.5555/2984093.2984314 (Accessed: 21 July 2024).

Wang, X., & McCallum, A. (2006). Topics over Time. A Non-Markov Continuous-Time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 424–433). New York: Association for Computing Machinery. DOI: https://doi.org/10.1145/1150402.1150450 (Accessed: 21 July 2024).

Wehrheim, L. (2019). Economic History Goes Digital. Topic Modeling the Journal of Economic History, *Cliometrica,* 13(1), 83–125. DOI: https://doi.org/10.1007/s11698-018-0171-7 (Accessed: 18 June 2024).

Id., Spoerer, M., & Jopp, T.A. (2022). Turn, Turn, Turn. A Digital History of German Historiography, 1950–2019, *The Journal of Interdisciplinary History,* 53(3), 471–507. DOI: https://doi.org/10.1162/jinh_a_01871 (Accessed: 18 June 2024).

Zhao, X., Wang, D., Zhao, Z., Liu, W., Lu, Ch., & Zhuang, F. (2021). A Neural Topic Model with Word Vectors and Entity Vectors for Short Texts, *Information Processing & Management,* 58(2), 1–11. DOI: https://doi.org/10.1016/j.ipm.2020.102455 (Accessed: 18 June 2024).

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., & Xiong, H. (2016). Topic Modeling of Short Texts. A Pseudo-Document View. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2105–2114). New York: Association for Computing Machinery. DOI: https://doi.org/10.1145/2939672.2939880 (Accessed: 18 June 2024).

Figure Credits

All figures were created by the author herself. Fig. 1–3 were generated with the Gensim wrapper for *MALLET* and visualised with the *Python* library *WordCloud*[23]. Fig. 4–6 were created with *Miro*[24].

Fig. 5f. are based on Althage (2023). All other figures were first published here.

23 See https://amueller.github.io/word_cloud (Accessed: 18 June 2024).
24 See https://miro.com/de (Accessed: 18 June 2024).

Sentiment Analysis

Rachele Sprugnoli

https://orcid.org/0000-0001-6861-5595

Abstract This chapter presents an overview of Sentiment Analysis with a focus on how it is typically applied in the Digital Humanities field. More specifically, we discuss linguistic issues, such as irony and the use of emoji, that make sentiment analysis challenging and we provide a brief description of several tasks and sub-tasks, all related to subjective texts but seen from different angles: i.e., subjectivity classification, document- and sentence-level polarity classification, aspect-based sentiment analysis, stance detection, irony detection and emotion analysis. In addition, we introduce lexicon-based and machine learning approaches to sentiment analysis. Open issues and best practices for the application of sentiment analysis methods in Digital Humanities are also discussed and the chapter closes with a list of emergent trends in the field.

Keywords Sentiment Analysis, Opinion Mining, Emotion Analysis

1. Introduction

Sentiment Analysis (SA) is a field of research, within the area of Natural Language Processing (NLP),[1] that aim to identify and classify opinions, feelings, personal evaluations towards entities (e.g., people, places, products), events, topics as expressed in written texts (Liu 2022). In its simplest form, SA distinguishes texts according to their polarity (or sentiment orientation): "I love detective stories" has a positive polarity, "I don't like romance books" has a negative polarity and "Agatha Christie was an English writer" has a neutral polarity.[2]

There are many alternative expressions used in the literature to refer to this multifaceted problem: we find, among others, opinion mining, opinion extraction, sentiment mining, affect analysis, polarity detection. In this context the words *sentiment* and *opinion* are often considered synonyms; although they are not, their distinction is very subtle, and they are closely connected. The sentence "I get bored reading

1 A distinction is traditionally made between Computational Linguistics, seen as a branch of linguistics, and Natural Language Processing, seen as a branch of engineering or computer science (Bender 2016). However, in this chapter, we will adopt an integrated view of these two areas since both have the goal of carrying out linguistic analysis and use linguistic data as input.

2 Unless otherwise specified, the examples in this chapter were created by the author.

Rachele Sprugnoli: Sentiment Analysis, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 267–282. DOI: https://doi.org/10.11588/heibooks.1521.c21953

romantic books" expresses a negative sentiment prompted by the feeling of boredom, whereas "I think romantic books all have the same plot" expresses a negative opinion; this example shows that a negative sentiment implies a negative opinion, and a negative opinion is due to a negative sentiment. SA also includes other areas of research and applications that require more granular distinctions which will be addressed in a specific section; for example, aspect-based SA identifies the sentiment of specific attributes or components of an entity.

The growth of interest in SA goes hand in hand with the increasing diffusion of online reviews, forums, microblogs, and social networks which produce an enormous volume of subjective texts, in which users express their opinions and evaluations. SA is also considered a valid tool in the corporate, communicative, and social science fields: in fact, there are many applications which monitor the opinion of customers towards a service or product, or which study the attitude of users on social networks. There are also works that adapt methods and techniques of SA to the humanities with applications to historical, literary, or classical language texts.

In this chapter we introduce the basic definitions and concepts related to SA research with the aim of making the reader aware of the challenges related to SA, especially in the field of Digital Humanities (DH).

2. Why Sentiment Analysis is Challenging – Some Linguistic Peculiarities

The examples given in the previous section are extremely simple from a linguistic point of view but the language we use to express our subjective evaluations is complex, made up of many components that make SA an interesting challenge both for humans and computers.

First of all, the same word in different contexts can have different meanings that encode different sentiments. For example, the adjective *sharp* can be associated with a negative sentiment when it means "keenly and painfully felt" but has a positive sentiment when it means "having or demonstrating ability to recognize or draw fine distinctions."[3]

Furthermore, opinions are not always expressed explicitly and directly but often have an implicit or comparative form. Implicit opinions are those referring to facts or effects related to the object of the opinion: for example, the sentence "this book just makes me yawn" describes a side effect of reading a boring book. Comparative opinions, on the other hand, juxtapose different elements based on the same aspect as in

3 Definitions taken from WordNet 3.1: http://wordnetweb.princeton.edu/perl/webwn (Accessed: 23 June 2024).

"I think Agatha Christie's novels have a more linear plot than those of Arthur Conan Doyle:" it is interesting to note that understanding the sentiment of this last example is difficult because it depends on the reader's personal taste.

Implicit opinions often require world (extra-linguistic) knowledge to be correctly interpreted. The sentence "She looks like a Botticelli madonna!" expresses a positive sentiment by referring to the harmony and beauty of the faces painted by the Renaissance artist; on the contrary, "He looks like a Picasso painting!" makes a reference to the unstructured faces of cubism and therefore to a face with disproportionate features.

World knowledge is also needed to discriminate literal from ironic content. Irony is a type of figurative language that is intentionally used to give a sentence an opposite meaning to the literal one. As defined by Utsumi (2000) in his *Implicit Display Theory,* verbal irony is an utterance or a statement that implicitly displays an ironic environment in which the speaker has a negative emotional attitude toward the incongruity between what he/she expects and what actually is. The term irony is often used as a hypernym for sarcasm (Grice 1975) that indeed is a particular form of irony used to mock or insult in a scornful or caustic way. Both irony and sarcasm are particularly interesting in SA because they are sentiment shifters, i.e., they change the polarity: a sentence like "the wifi connection is great – it's fast as a sloth" means the exact opposite of what it seems (the wifi connection is slow) but its apparent linguistic form would lead to assign it a positive polarity.

Another issue to consider is the presence of emoticons and emojis that play an important role when dealing with informal texts such as posts on social networks and forums. Comparing "Rome :)" to "Rome :(," the opinion on the city is expressed by the emoticon; without it, the name Rome alone would have no polarity. In other words, these elements enhance the expressiveness of a text and convey their own specific sentiment even if not always easily identifiable. For example, the fire emoji is mostly used with the meaning of excellent or attractive (therefore with a positive sentiment), but can also signal anger (thus a negative sentiment) or a fact, such as the presence of fires or excessive heat (having, in this case, a neutral sentiment).

3. One Name, Many Tasks

As already stated by Liu in 2010, SA is a multifaceted problem: it is not a single monolithic linguistic task, it does not have a single solution but can be tackled by considering various levels of analysis.

The first level is addressed by the task called subjectivity classification which aims to distinguish objective texts, containing factual information, from subjective, opinionated texts that express feelings, points of view or personal beliefs. This is the first step towards more in-depth analyses: in fact, in objective texts it is not possible to identify a polarity (they are neutral) while subjective texts can be classified according to their sentiment orientation.

Polarity classification is the next step and consists of assigning to an information unit a value that indicates whether it expresses a positive, negative, or neutral sentiment. This value can be categorical or numeric and the range of possible values can vary considerably depending on the degree of detail we want to achieve. For example, there are binary classifications (with only two values, such as *positive* and *negative*), 3-value classifications (e.g., *positive, neutral, negative* or *+1, 0, -1*), 5-value classifications (e.g., *very positive, somewhat positive, neutral, somewhat negative, very negative* or *+1, +0.5, 0, -0.5, -1*) but also decimal scores in a continuous range (typically between *+1* and *-1*).

Polarity classification can be performed at different granularities, i.e., taking into consideration different types of information units: the whole document, a single sentence at a time, or one specific aspect. Document level SA assigns a polarity score to an entire document (e.g., a book review) by assuming such document as a single information unit expressing the opinion of a single person (the author of the review) on a single entity (a book). The same type of classification can be applied at sentence level. Sentence-level SA is useful because the same document can contain different or even opposite opinions in different sentences. For example, a book review may be made up of neutral sentences, describing the plot without making personal judgments, together with other sentences expressing appreciation or disapproval. An even more granular level of analysis is provided by the entity-based or aspect-based SA,[4] which has the purpose of extracting the opinions expressed on individual entities or on entities' features. In the case of the aforementioned review, the book is the entity object of the evaluation while two relevant features can be the plot and the price; the sentiment can be different for each of these elements, for example it can be positive for the book itself and for the plot but negative for the price as in "I enjoyed reading the book because the storyline is compelling, but the price is too high: not everyone can spend 25 euros on a book!" Therefore, the task has two main phases: the extraction of entities and/or features and then the classification of the sentiment for each of them. It is important to note that the relevant features are entity type specific: if price is important to any commercial product or service, plot is specific to books and movies. A mobile phone, on the other hand, can have battery life and ease of configuration as features to identify, while for hotels the location is particularly important.

4 Aspect-based SA is also known as feature-based SA.

4. Other Related Tasks

In this section, we provide an overview of other tasks that are considered sub-problems of SA, all related to subjective texts but seen from different angles.

- *Stance detection* is the task that determines whether the author of a text is in favor or against an entity, event, or topic (AlDayel & Magdy 2021). From a linguistic point of view, stance is an overt expression used to evaluate a certain target element and position oneself with respect to the others by displaying alignment or opposition (Du Bois 2007). For this reason, stance detection requires a given target to measure the author's viewpoint toward it and the output of the classification is one out of the three labels *Favor, Against, Neither,* instead of *Positive, Negative* or *Neutral* as in the simplest case of polarity classification. Stance and polarity are independent of each other: a positive sentiment does not necessarily lead to a supporting stance, just as a negative sentiment is not necessarily associated with an opposing stance. For example, taking the statement "climate change is a real concern" as target, the sentence "It's so sad that too many people don't plan to do anything while our planet is burning!" expresses a negative sentiment but a supportive stance towards the statement. This task is mostly applied to political and social issues to intercept the position of social network users regarding a political figure or proposals considered divisive, such as drug liberalization and same-sex marriage.
- *Irony detection and sarcasm detection* tasks aim to distinguish between ironic or sarcastic and non-ironic or non-sarcastic texts (Maynard & Greenwood 2014). While irony is usually uncritical, sarcasm is more aggressive; however, both these figurative devices create a mismatch between the literal and the intentional meaning of a text. Sometimes a binary classification is made without differentiating between irony and sarcasm, while in other cases a more detailed classification is attempted by recognizing various types of irony, for example, by distinguishing it from sarcasm, satire or parody (Abu Farha et al. 2022).
- *Emotion analysis* consists of determining which emotions are conveyed in a text. The scientific study of emotions has interested psychologists and anthropologists since the publication of Darwin's seminal work *The Expression of the Emotions in Man and Animals* in 1872. Although the theories are numerous, there are two main approaches on which computational techniques are based. According to the first approach, emotions are innate, universal across different cultures and limited in number thus they can be classified using categorical labels. In NLP, these labels are often borrowed from the theories defined by Ekman (1993) or Plutchik (1980). Ekman identifies six emotions (i.e., anger, disgust, sadness, joy, fear, surprise) while Plutchik defines four

pairs of basic emotions (i.e., joy versus sadness, anger versus fear, trust versus disgust, surprise versus anticipation) which combine with each other to form dyads, i.e., complex emotions (e.g., love is a combination of joy and trust). On the other hand, following the second approach, emotions cannot be labeled but represented according to different dimensions using continuous values. In the circumplex model (Russell 1980) there are two fundamental dimensions of emotional experience: valence, i.e., the level of pleasantness, and arousal, i.e., the intensity of the emotion. A third dimension called dominance is often added to the previous two to encode the degree of control the emotion exerts over the person experiencing it. For example, according to this approach known with the acronym VAD *(Valence-Arousal-Dominance),* anger has low valence, high arousal, and high dominance.

5. Methods

From the point of view of the development of SA systems, two main approaches can be distinguished: those based on lexicons, and those using machine learning algorithms, both supervised and unsupervised.

Lexicon-based methods rely on the intuition that the polarity of a text can be obtained on the basis of the polarity of the words that compose it (Taboada 2011). Such polarity is obtained from lexicons made up of lists of tokens, lemmas, or phrases in which each lexical entry is associated with a categorical or numerical value (e.g., *Positive* or *+1*) quantifying its sentiment orientation. Polarity lexicons are available for numerous languages (Mohammed & Balakrishnan 2020): some have been created manually, employing experts (e.g., linguists or psychologists) or crowdsourcing techniques (Mohammad & Turney 2013),[5] but the development of these resources is very time-consuming thus automatic approaches have also been tested for example by exploiting machine translation or available lexicographical resources and corpora. Lexicons typically record the prior polarity of words, i.e., the sentiment they evoke beyond their context of use. Thus, words like *friendship* and *love* are associated with a positive polarity while *murder* and *hate* with a negative one. Rarer are the lexicons that contain sense-based polarities, the best known is *SentiWordNet* (Baccianella et al. 2010) in which each *WordNet synset* (Miller 1995) has a positive, a negative and an objective score. Based on these lexicons, scripts are created which calculate the ratio between positive and negative words within the text to be analyzed: if the text has more positive words it is classified as positive, otherwise it is classified as negative. This approach is very simple to apply but tends to be less accurate than machine learning methods because the lexicon coverage is not unlimited and because the specific

5 The work is carried out by non-expert collaborators recruited on specific web platforms.

SenticNet

POLARITY
positive

NRC-VAD-Lexicon

VALENCE	AROUSAL	DOMINANCE
0.802	0.549	0.647

SentiWordNet 3.0

PoS	Synset ID	PosScore	NegScore	Gloss
n	07186148	0	0	a request (spoken or written) to participate…
n	04689048	0.5	0	a tempting allurement

DepecheMood++

AFRAID	AMUSED	ANGRY	ANNOYED	DONT_CARE	HAPPY	INSPIRED	SAD
0.045481	0.155150	0.130596	0.154374	0.179065	0.135442	0.148765	0.051123

NRC-Emotion-Lexicon

ANGER	ANTICIPATION	DISGUST	FEAR	JOY	SADNESS	SURPRISED	TRUST	POSITIVE	NEGATIVE
0	1	0	0	0	0	0	0	1	0

Fig. 1 Entries for the noun *invitation* in different polarity and emotion lexicons.

context can vary the polarity of a word. It is important to note that the lexicon-based approach can be also applied in the emotion analysis task: in this case emotion lexicons containing word-emotion associations are used. Fig. 1 shows how different 5 English lexicons are from each other in the way they assign polarity or emotional values to the same word i.e., the noun *invitation.* The lexicons taken into consideration for this comparison are: *SenticNet* (Cambria et al. 2022), *NRC-VAD-Lexicon* (Mohammad 2018), *SentiWordNet 3.0, DepecheMood++* (Araque et al. 2019), *NRC-Emotion-Lexicon* (Mohammad & Turney 2013).

Machine learning, in the context of NLP, is the process of training a computational system to perform a certain linguistic task. In the supervised approach, the algorithm is trained taking as input a set of annotated data, that is a selection of texts in which the expected classification is provided (for example a collection of sentences each associated with a polarity value). On the contrary, in the unsupervised method training data is not provided but the system tries to autonomously extract generalizations from input texts. Generally, unsupervised learning tends to be less expensive than supervised learning, as it does not require training data, but the results are less accurate. For this reason, there are numerous initiatives that aim to produce annotated data for all the tasks mentioned in the previous sections, covering many languages and various textual genres. Over time, the machine learning algorithms used have evolved and deep learning techniques are now more widely adopted, leading to major improvements in system performance (Yadav & Vishwakarma 2020).

Whichever method is used, system performances tend to vary greatly depending on the task, text types and granularity of analysis. In general, the greater the number of labels used for classification, the greater the complexity of the task and therefore the lower the performance (Wankhade 2022).

6. Sentiment Analysis in Digital Humanities Research

Although the most common resources and tools for SA fall into categories such as social network analysis and customer opinion monitoring, research in DH has been growing in recent years. In general, the interest in the use of NLP methods for the processing of humanistic data is rising, as demonstrated by the large participation in dedicated scientific events.[6] In this increasingly rich panorama of projects and activities at the intersection between DH and NLP, SA is considered a fruitful technique for enriching textual data especially in the fields of history and literary studies.

Most works in the historical domain primarily use digitized newspaper articles as data to understand how important events or famous figures were perceived by their contemporaries. For example, entity-based SA is used in the *Oceanic Exchanges* project to identify the opinion expressed in 19th century German newspapers towards a group of writers of the same period (Keck et al. 2020), while Viola (2023) employs the same method for analyzing the sentiment towards a selection of entities in US newspapers published by Italian immigrants. On the other hand, Mayer et al. (2022) studies the transnational reception of the execution of Maximilian I, emperor of Mexico, in 1867 relying on newspaper from various countries. The case study presented by Sprugnoli et al. (2016) is different because it detects both prior and contextual polarities in Italian political texts of the first half of the 20th century and demonstrates that sentiment orientation is often implicitly expressed, making it particularly difficult to assign a polarity value even for humans.

The spectrum of research in the field of computational literary studies is broader. Starting from the pioneering work of Anderson & McMaster (1982) on the measurement of affective tones in the chapter of a novel and in a set of children's stories, the applications concerns various textual genres (gothic and romantic novels, fairy tales, plays, fan fiction) and various purposes (understanding what makes one plot more intriguing than another, what role emotions play in the interactions between characters, how emotions can help distinguish between different literary genres, what are

6 See, e.g., the annual workshops of the *ACL Special Interest Group on Language Technologies for the Socio-Economic Sciences and Humanities* (LaTeCH, https://sighum.wordpress.com/events/), the *Computational Humanities Research* (CHR, https://2023.computational-humanities-research.org) Conference, the *Workshop on Ancient Language Processing* (ALP, https://www.ancientnlp.com/alp2023) and the *Workshop on Language Technologies for Historical and Ancient Languages* (LT4HALA, https://circse.github.io/LT4HALA). All addresses were accessed on 23 June 2024.

the emotional arcs of stories) as well described in the survey papers by Kim & Klinger (2019) and Rebora (2023) to which we refer for further details.

7. Open Issues and Best Practices

The works cited so far (as well as all those that we have not been able to cite due to space limitations) have had, in one way or another, to address various issues relating to research practices in the humanities and to the characteristics of humanistic texts written in non-contemporary languages. First, literary, and historical texts are often sparse, inconsistent and incomplete, presenting many orthographic variations due to diachrony and diatopy phenomena. Then, to be processed by NLP systems, texts must be available in machine-readable format: the use of OCR *(Optical Character Recognition)* systems in digitization processes, especially when applied to manuscripts or ancient prints, is not free from errors and it is often necessary to intervene to reduce noise and obtain high quality data. Furthermore, humanities scholars work on textual genres (such as poems, plays, philosophical and historical treatises) that are very different from those usually analyzed by NLP systems: this requires that such systems be appropriately adapted or developed from scratch. Finally, final users of DH applications are humanities scholars who are often not tech-savvy users, so it is important to develop simple, intuitive and transparent systems.

The lack of large amounts of data on which to train machine learning systems, combined with the demand for systems whose results are easily interpretable has led to the widespread adoption of the lexicon-based approach in DH (Ohman 2021). In fact, machine learning algorithms are often criticized because they are difficult to interpret; they are like black boxes and not even the developers are able to explain perfectly why certain choices and, consequently, certain predictions are made. On the contrary, lexicon-based systems make it easier to understand the results, to highlight trends and passages which can be then re-analyzed through a closer reading. Furthermore, the need for intuitive systems has result in the creation of user-friendly graphical interfaces, more suitable for use by non-experts than programming scripts; some examples are *SEANCE* (Crossley et al. 2017), *Lingmotif* (Moreno-Ortiz 2017), and *SentText* (Schmidt et al. 2021). It is important to notice that *Syuzhet,* the first SA system that had a notable resonance, but also numerous criticisms, in the DH community, is lexicon-based and extremely straightforward from a computational point of view since it is based on simple word count;[7] since then, however, lexicon-based approaches have become more refined and the aforementioned tools include preprocessing functionalities (e.g., stop-word removal, lemmatization) and rules to handle negations.

7 See http://www.matthewjockers.net/2015/02/02/syuzhet (Accessed: 23 June 2024).

Whichever method is used and whether a polarity or emotion analysis is to be performed, there are various aspects to consider (Mohammad 2023). In fact, it is necessary to choose lexicon, data, and system to use, or decide to develop new ad hoc resources suitable for the domain of interest because the existing ones are not in line with the objectives of the research. This involves choosing the type of conceptualization to adhere to, i.e., whether to opt for the categorical or dimensional approach, but also whether to use continuous values or discrete labels, as well as the best level of granularity (in other words, how many classes or how many dimensions you want to capture). In addition, when a new lexicon or a new annotated dataset is to be developed from scratch, it's crucial to choose whether to recruit expert or non-expert annotators, through the adoption of crowdsourcing techniques. This second option, although widely employed when dealing with texts from social networks and contemporary languages, is more difficult to apply when dealing with historical and ancient languages. Furthermore, specifically in the case of ancient languages, the problem of the lack of native speakers must also be addressed because it is impossible to rely on the intuition or personal sensitivity of the annotators, thus it is essential to involve language and culture experts (Sprugnoli et al. 2020).

Defining the most correct procedure to follow can be a long interactive process, made up of several experimentation phases. For example, Schmidt et al. (2021) details the choices made to define a new scheme for annotating emotions in German plays written around 1800. Although at first, they considered adopting the categorical approach using basic emotions defined by Ekman or Plutchik, they soon realized that the psychological theories on which these categories were based did not reflect emotion and affect concepts of literary theories. Thanks to a pilot annotation, they noticed that some emotions were particularly relevant even if they did not belong to any psychological theory (e.g., friendship), while others did not have great importance (e.g., disgust) in dramatic texts. In the end, they came up with a new hierarchical scheme made up of 13 emotion concepts. Another interesting example is given by the analysis of emotions in poems which shows how the same textual genre can be addressed by considering different aspects (Sprugnoli et al. 2023) such as the level of expertise of annotators (experts or crowd workers), the textual unit to be annotated (line, sentence, stanza, whole poem), the number of emotions considered (two or more), the general perspective (emotions are annotated as intended by the author or as perceived by the reader). For instance, each poem of PO-EMO is annotated at both line and stanza levels with 9 emotions elicited in the reader employing both trained experts and crowd workers (Haider et al. 2020). On the other hand, only two classes are assigned by experts at each poem in the Kabithaa corpus made of Odia poems (Mohanty et al. 2018).

Summing up, there is no single right way to proceed: all the choices must be weighted based on the research objective, the characteristics of the texts to be analyzed, and the theoretical context of reference.

8. Conclusion

This chapter described the complexity of SA seen as a multifaceted problem: different tasks, methods and applications are introduced by focusing on research in the DH field. To conclude, we mention five promising lines of research that have emerged in recent years attracting increasing attention.

Large Language Models (LLMs). LLMs are deep neural networks (called transformers) trained on massive amounts of unannotated data to predict how a sentence continues or what is missing in a sentence. Prompts, that is instructions in natural language describing the task to perform, are the way humans interact with a LLM.[8] LLMs displays impressive capacities in many NLP tasks but understanding how to optimize prompts to achieve increasingly better results is an open issue also in the case of SA (Mao et al. 2023).

Multimodality. Multimodal SA allows to go beyond text-based SA integrating linguistic information with audio-visual information extracted from images, audio recordings and videos. The first experiments in the DH field concerned the analysis of plays (Schmidt & Wolff, 2021) and oral history interviews (Gref et al., 2022).

Linguistic Linked Data. Linguistic resources (lexicons and annotated data) for SA are now very numerous but do not interact with each other: using linked data techniques (Iglesias et al. 2017) would make them interoperable, more visible and reusable. Data models, ontologies and interlinked resources are presented every year during the Sentiment Analysis & Linguistic Linked Data workshop series, also including papers on Classical languages (Sprugnoli et al. 2021).

Perspectivism. To train machine learning systems it is necessary to have high quality data, in which labels are assigned in a consistent way; however, annotation is highly subjective when it comes to sentiment and emotions and it is often difficult to have a consensus on the label to assign because multiple interpretations are possible, especially when dealing with literary texts. In case of doubt or disagreement the assignments are forced towards a single label so that the algorithm can learn and make predictions. To change this paradigm, the so-called perspectivism, a more inclusive framework that aim at preserving the different points of view of the annotators, has been proposed (Cabitza et al. 2023).

Reader Response Studies. If the application of SA to the analysis of literary texts is still subject to criticism because it is not easy to find the right balance between computational approaches and narratological theories, reader response studies are

8 An example of such interaction is given by the ChatGPT interface.

finding greater success also in terms of system performance, showing that SA seems to be more efficient when applied to comments on a literary text than on the text itself (Pianzola et al. 2020).

References

Abu Farha, I., Oprea, S.V., Wilson, S., & Magdy, W. (2022). SemEval-2022 Task 6. iSarcasmEval. Intended Sarcasm Detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation* (SemEval-2022) (pp. 802–814). Seattle: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2022.semeval-1.111 (Accessed: 23 June 2024).

AlDayel, A., & Magdy, W. (2021). Stance detection on social media. State of the art and trends, *Information Processing & Management,* 58(4), 1–22.

Anderson, Clifford W., & McMaster, G.E. (1982). Computer assisted modeling of affective tone in written documents. *Computers and the Humanities,* 16(1), 1–9.

Araque, O., Gatti, L., Staiano, J., & Guerini, M. (2019). Depechemood++. A bilingual emotion lexicon built through simple yet powerful techniques, *IEEE transactions on affective computing,* 13(1), 496–507. DOI: https://doi.org/10.1109/TAFFC.2019.2934444 (Accessed: 23 June 2024).

Bender, E.M. (2016). Linguistic typology in natural language processing, *Linguistic Typology,* 20(3), 645–660.

Cabitza, F., Campagner, A., & Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing, *Proceedings of the AAAI Conference on Artificial Intelligence,* 37(6), 6860–6868. DOI: https://doi.org/10.1609/aaai.v37i6.25840 (Accessed: 23 June 2024).

Cambria, E., Liu, Q., Decherchi, S., Xing, F., & Kwok, K. (2022). SenticNet 7. A commonsense-based neurosymbolic AI framework for explainable sentiment analysis, *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 3829–3839. URL: https://aclanthology.org/2022.lrec-1.408 (Accessed: 23 June 2024).

Crossley, S.A., Kyle, K., & McNamara, D.S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE). An automatic tool for sentiment, social cognition, and social-order analysis, *Behavior research methods,* 49, 803–821.

Darwin, Ch. (1872). *The Expression of the Emotions in Man and Animals.* London: John Murray.

Du Bois, J.W. (2007). The stance triangle. In R. Engebretson (Ed.). *Stancetaking in Discourse. Subjectivity, evaluation, interaction* (pp. 139–182). Amsterdam: John Benjamins.

Ekman, P. (1993). Facial expression and emotion, *American psychologist,* 48(4), 384–392.

Gref, M, Matthiesen, N., Venugopala, S.H., Satheesh, Sh., Vijayananth, A., Ha, D.B., Behnke, S., & Köhler, J. (2022). A Study on the Ambiguity in Human Annotation of German Oral History Interviews for Perceived Emotion Recognition and Sentiment Analysis, *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 2022–2031. URL: https://aclanthology.org/2022.lrec-1.217 (Accessed: 23 June 2024).

Grice, H.P. (1975). Logic and conversation. In P. Cole & J.L. Morgan (Eds.), *Syntax and semantics,* 3. Speech acts (pp. 41–58). New York/San Francisco/London: Academic Press.

Haider, Th., Eger, S., Kim, E., Klinger, R., & Menninghaus, W. (2020). PO-EMO. Conceptualization, Annotation, and Modeling of Aesthetic Emotions in German and English Poetry, *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 1652–1663. URL: https://aclanthology.org/2020.lrec-1.205 (Accessed: 23 June 2024).

Hernández Farias, D.I., & Rosso, P. (2017). Irony, sarcasm, and sentiment analysis. In F.A. Pozzi, E. Fersini, E. Messina & B. Liu (Eds.), *Sentiment Analysis in Social Networks* (pp. 113–128). Amsterdam et al.: Morgan Kaufmann.

Hutto, C., & Gilbert, E. (2024). VADER. A parsimonious rule-based model for sentiment analysis of social media text, *Proceedings of the international AAAI conference on web and social media,* 8(1), 216–225. DOI: https://doi.org/10.1609/icwsm.v8i1.14550 (Accessed: 23 June 2024).

Iglesias, C.A., Sanchez-Rada, F.J., Vulcu, G., & Buitelaar, P. (2017). Linked data models for sentiment and emotion analysis in social networks. In F.A. Pozzi, E. Fersini, E. Messina & B. Liu (Eds.), *Sentiment Analysis in Social Networks* (pp. 49–69). Amsterdam et al.: Morgan Kaufmann.

Keck, J., Knabben, M., & Pado, S. (2020). Who's in the News? Methodological Challenges and Opportunities in Studying 19th-century Writers in Historical Newspapers. *Europeana PRO,* 16. Newspapers, no pag. URL: https://pro.europeana.eu/page/issue-16-newspapers#who-s-in-the-news (Accessed: 23 June 2024).

Kim, E., & Klinger, R. (2019). A Survey on Sentiment and Emotion Analysis for Computational Literary Studies, *Zeitschrift für digitale Geisteswissenschaften,* no pag. DOI: https://doi.org/10.17175/2019_008_v2 (Accessed: 23 June 2024).

Liu, B. (2010). Sentiment analysis. A multi-faceted problem. *IEEE intelligent systems,* 25(3), 76–80.

Id. (2022). *Sentiment analysis and opinion mining.* Cham: Springer. DOI: https://doi.org/10.1007/978-3-031-02145-9 (Accessed: 23 June 2024).

Mayer, A.I.L., Gutierrez-Vasques, X., Saiso, E.P., & Salmi, H. (2022). Underlying Sentiments in 1867. A Study of News Flows on the Execution of Emperor Maximilian of Mexico in Digitized Newspaper Corpora, *Digital Humanities Quarterly,* 16(4), 1–98. URL: http://www.digitalhumanities.org/dhq/vol/16/4/000649/000649.html (Accessed: 23 June 2024).

Maynard, D., & Greenwood, M. (2014). Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis, *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), 4238–4243. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/67_Paper.pdf (Accessed: 23 June 2024).

Mao, R., Liu, Q., He, K., Li, W., & Cambria, E. (2023). The biases of pre-trained language models. An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3), 1743–1753. DOI: https://doi.org/10.1109/TAFFC.2022.3204972 (Accessed: 23 June 2024).

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words, *Proceedings of the 56th annual meeting of the association for computational linguistics,* 1. Long papers, 174–184. DOI: https://doi.org/10.18653/v1/P18-1017 (Accessed: 23 June 2024).

Id. (2023). Best Practices in the Creation and Use of Emotion Lexicons, *Findings of the Association for Computational Linguistics.* EACL 2023, 1825–1836. DOI: https://doi.org/10.18653/v1/2023.findings-eacl.136 (Accessed: 23 June 2024).

Id., & Turney, P.D. (2013). Crowdsourcing a word-emotion association lexicon, *Computational intelligence,* 29(3), 436–465. DOI: https://doi.org/10.1111/j.1467-8640.2012.00460.x (Accessed: 23 June 2024).

Mohanty, G., Mishra, P., & Mamidi, R. (2018). Kabithaa. An annotated corpus of Odia poems with sentiment polarity information, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), 52–57. URL: http://lrec-conf.org/workshops/lrec2018/W11/pdf/15_W11.pdf (Accessed: 23 June 2024).

Moreno-Ortiz, A. (2017). Lingmotif. Sentiment analysis for the digital humanities, *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics,* 73–76. URL: https://aclanthology.org/E17-3019 (Accessed: 23 June 2024).

Ohman, E. (2021). The validity of lexicon-based emotion analysis in interdisciplinary research, *Proceedings of the Workshop on Natural Language Processing for Digital Humanities* (NLP4DH), 7–12. URL: https://aclanthology.org/2021.nlp4dh-1.2 (Accessed: 23 June 2024).

Pianzola, F., Rebora, S., & Lauer, G. (2020). Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins, *PloS one,* 15(1), 1–46. URL: 10.1371/journal.pone.0226708 (Accessed: 23 June 2024).

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In id. & H. Kellermann (Eds.), *Theories of emotion* (pp. 3–33). Cambridge, Mass.: Academic press [= *Emotion. Theory, Research, and Experience,* 1]. DOI: https://doi.org/10.1016/B978-0-12-558701-3.50007-7 (Accessed: 23 June 2024).

Rebora, S. (2023). Sentiment Analysis in Literary Studies. A Critical Survey, *Digital Humanities Quarterly,* 17(2), 1–50. URL: http://www.digitalhumanities.org/dhq/vol/17/2/000691/000691.html (Accessed: 23 June 2024).
Russell, J. A. (1980). A circumplex model of affect, *Journal of Personality and Social Psychology,* 39, 1161–1178.
Schmidt, Th, & Wolff, Ch. (2021). Exploring multimodal sentiment analysis in plays. A case study for a theater recording of Emilia Galotti, *Proceedings of CHR 2021. Computational Humanities Research Conference,* 392–404. URL: https://ceur-ws.org/Vol-2989/short_paper45.pdf (Accessed: 23 June 2024).
Schmidt, Th., Dangel, J., & Wolff, Ch. (2021). A Tool for Lexicon-based Sentiment Analysis in Digital Humanities, *Proceedings of the 16th International Symposium of Information Science* (ISI 2021), 156–172. URL: https://epub.uni-regensburg.de/44943/1/isi_schmidt_dangel_wolff.pdf (Accessed: 23 June 2024).
Schmidt, Th., Dennerlein, K., & Wolff, Ch. (2021). Towards a corpus of historical german plays with emotion annotations, *Proceedings of 3rd Conference on Language. Data and Knowledge (LDK 2021). Schloss Dagstuhl. Leibniz-Zentrum für Informatik,* 1–11 [= *Open Access Series in Informatics,* 93]. DOI: https://doi.org/10.4230/OASIcs.LDK.2021.9 (Accessed: 23 June 2024).
Sprugnoli, R., Passarotti, M., Corbetta, D., &. Peverelli, A. (2020). Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin, *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 3078–3086. URL: https://aclanthology.org/2020.lrec-1.376 (Accessed: 23 June 2024).
Sprugnoli, R., Passarotti, M., Testori, M., & Moretti, G. (2022). Extending and Using a Sentiment Lexicon for Latin in a Linked Data Framework, *Proceedings of the Workshops and Tutorials. Language Data and Knowledge 2021 (LDK 2021),* 1–14. DOI: https://doi.org/10.5281/zenodo.6303164 (Accessed: 23 June 2024).
Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis, *Computational linguistics,* 37(2), 267–307. DOI: https://doi.org/10.1162/COLI_a_00049 (Accessed: 23 June 2024).
Utsumiu, A. (2000). Verbal irony as implicit display of ironic environment. Distinguishing ironic utterances from nonirony, *Journal of pragmatics,* 32(12), 1777–1806. URL: http://www.utm.se.uec.ac.jp/~utsumi/paper/jop2000-utsumi.pdf (Accessed: 23 June 2024).
Viola, L. (2023). Networks of migrants' narratives. A post-authentic approach to heritage visualisation, *ACM Journal on Computing and Cultural Heritage,* 16(1), 1–21. DOI: https://doi.org/10.1145/3575863 (Accessed: 23 June 2024).
Wankhade, M., Rao, A. Ch. S., & Kulkarni, Ch. (2022). A survey on sentiment analysis methods, applications, and challenges, *Artificial Intelligence Review,* 55(7), 5731–5780. DOI: https://doi.org/10.1007/s10462-022-10144-1 (Accessed: 23 June 2024).
Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures. a review, *Artificial Intelligence Review,* 53(6), 4335–4385. DOI: https://doi.org/10.1007/s10462-019-09794-5 (Accessed: 23 June 2024).

Figure Credit

Fig. 1 was created by the author herself and first published here.

Intertextuality Research

Julia Nantke

https://orcid.org/0000-0002-8037-1009

Abstract The concept of intertextuality refers to the relationships, functions, and effects arising between two or more texts where the texts reference each other in quotations, allusions, or structural parallels. The concept was formulated in literary studies in the 1970s and 1980s as part of an extensive theoretical debate and has recently been updated under the auspices of digitality. Digital methods are used to find, annotate, and evaluate intertextual references. Depending on the method, different approaches to the phenomenon of intertextual relationships develop, some of which align with traditional literary studies concepts and are characterized by the specifics of digitality.*

Keywords Intertextuality, Literary Theory, Text Reuse, Annotation, Operationalization, Modelling

The concept of intertextuality – i.e., the relationship between two or more texts marked by textual similarity – establishes a link between literary and theological research traditions. For (neo-philological) literary studies, the concept was established in the 1960s and 1970s by the post-structuralist conceptual and theoretical development surrounding Julia Kristeva and Roland Barthes. This concept was then further developed, specified, and partially reconceptualized in various ways during a subsequent phase (cf. Genette [1982] 1993; Riffaterre 1984; Broich & Pfister 1985; Lachmann 1990). By and large, the aim was to make the post-structuralist-deconstructivist concept of a text, which characterizes Kristeva's and Barthes' work, more operational by tracing it back to concretely verifiable text structures and/or authorial intention.[1] The embedding in the fundamental literary-theoretical debate on the meaning and function of central concepts and terms such as author, text, and work was decisive.

By way of contrast, the theological approach to textual similarity has been more practice oriented, in that the approach is less theoretical and more primarily describes

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

1 Cf. the following remarks on the intensity of intertextual markings in the compendium by Broich & Pfister: "Pretexts or text foils that are only brought to the text by the genesis of the work or only arbitrarily by the recipient constitute […] weak intertextual references, whereas the *hard center of maximum intensity* is reached when the *author* is *aware* of the intertextual reference, assuming that the recipient is also familiar with the *pretext* and refers to it clearly and unambiguously through a *conscious marking in the text*" (Pfister 1985, 27) [emphasis added, J.N.].

Julia Nantke: Intertextuality Research, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 283–293. DOI: https://doi.org/10.11588/heibooks.1521.c21954

the citation practices and textual references during the reception, dissemination, and transmission of biblical texts and theological literature (cf. Gillmayr-Bucher 2006). The reference to historical textual practices indicates the relevance of intertextual references, the historical tradition of which goes beyond the lines of the development of the concept of intertextuality to antiquity (cf. Berndt & Tonger-Erk 2013, 7). In part, the theological understanding, trained in a complex historical tradition with multi-layered, branched textual relationships, also comes much closer to the post-structuralist concept of a "mosaic of citations" (Kristeva [1967] 1972, 348) without a connection to a specific author than many of the literary scientific developments in scholarship (cf. also Brodie 2006, 75). At the same time, recent theological research also emphasizes the relevance of considering a *literary* dimension of biblical intertextuality (cf. Brodie et al. 2006, 4; id. 2006a, 285; Dörr 2012, 20–24). This study, in turn, explicitly corresponds with the literary tradition of intertextual analyzation.

Both disciplines have a long and diverse academic tradition of treating intertextuality phenomena in the age of (largely) analogous research in the humanities. These lines of tradition form the background against which the reconceptualization of the phenomenon of intertextuality by digital methods takes place. At the same time, many Digital Humanities methods are methodologically based on a comparison of textual properties,[2] which, however, takes place primarily based on numerical values and measurable characteristics on the surface of the text. Textual similarity therefore tends to be conceptualized differently in the context of computational operationalization than is the case for the classical humanities.

Therefore, mutual reference between theoretical conceptualization and method guided practice represents a central and, at the same time, flexible scope for digital intertextuality research, which is reflected in the development of different, partly overlapping approaches to researching intertextual phenomena. In the following, various approaches to the digital detection, modeling, and analysis of intertextual modes of writing are presented and compared in terms of their modeling practices, theoretical and methodological foundations. The relationship to analog traditions of intertextuality research are also considered.

1. Manual Digital Modeling of Intertextual Relationships

Systematically modeling intertextual references is one approach to increasing the possibilities for structuring and linking digital information for intertextual research. The starting point for this approach assumes intertextual references can be systematized regarding the specific relationship between the intertextually linked texts,

2 This assumption underlies the Working Group *Comparing Text* of the DFG Priority Program Computational Literary Studies led by Christof Schöch.

which already underlies the analog concepts of intertextuality. At the same time, digital implementation inductively gains new insight into the structure and functioning of intertextual relationships on the basis of a comprehensive machine-readable recording of intertextual references (cf. Nantke & Schlupkothen 2018; Horstmann et al. 2023). The conceptual point of reference is thus structuralist-hermeneutic theories of intertextuality, which, as shown in Genette (1993) or Broich & Pfister (1985), assume a clearly definable relationship between a source text and a subsequent text (cf. Molz 2020, 17). However, the digital no longer focuses on systematization based on a taxonomy of identifiable relationship types and degrees of intensity, but rather on the accumulation of data in a database or within the framework of a formal representation, which can then be evaluated according to common patterns (cf. Hohl-Trillini & Quassdorf 2010, 4; Nunn 2016). The aim of manual modeling approaches is to do justice to the complexity of the phenomenon of intertextuality by going beyond the *simple* cases of clearly identifiable, punctual references to include implicit references and structural parallels reaching beyond the individual text and affect the text. Additionally, the existence of an intertextual reference is modeled as well as how intertextual transformation and the functions of the intertextual references are also included (cf. Hohl-Trillini & Quassdorf 2010; Nantke & Schlupkothen 2018; Nantke & Schlupkothen 2019). A structured, machine-readable description of the intertextual relations then forms the basis for an automatic evaluation of the annotated features. In the most extensive project of this kind to date, *HyperHamlet,* a database created in 2010 combing over 8,000 references to Shakespeare's famous drama from literary and non-literary texts from the period 1600–2010. These texts are taken from the secondary literature on Hamlet and can be searched individually or together for various parameters as part of the digital presentation.[3] The *HyperHamlet* corpus also forms the basis for the WordWeb/IDEM database, which makes intertextual relationships between English dramas of the 16th and 17th centuries machine searchable and sortable.[4]

While the advantage of manual digital modeling of intertextuality lies primarily in the structured accumulation of data and its subsequent flexible evaluation, a manual approach also limits the scope of knowledge to the amount of data able to be generated. Approaches to collaborative annotation, such as those developed in the *TEASys* project for the annotation of intertextual relationships,[5] can extend this analyzable range of data.

A close reading approach, in which intertextual references are initially detected and annotated manually, is another starting point for automation. Three different

3 See http://www.hyperhamlet.unibas.ch (there was no access to the database on 27 July 2023. However, according to Regula Hohl-Trillini, the database will soon be accessible with this link).

4 See http://wordweb-idem.ch/index.html (Accessed: 18 June 2024).

5 Cf. http://www.annotating-literature.org/wp-content/uploads/2020/09/Styleguide-2020-08-11.pdf (Accessed: 18 June 2024).

options are conceivable. First, one can perform automated detection as a supplement to a qualitative-manual evaluation and then compare the results (cf. Molz 2020). Second, the manually recorded cases of intertextuality can be transferred into a formalized vocabulary, which then forms the basis for an automated evaluation that is potentially suitable for deriving new insights from the modelled data (cf. Nantke & Schlupkothen 2018; Horstmann et al. 2023). Third, under certain conditions, the data from manual annotation can be used as training data for an algorithm for the automated detection of intertextuality.[6]

Depending on which automation scenario is pursued, there are different requirements for the manually generated data. In his mixed methods study on Shakespearian references in works of contemporary British literature, Molz concentrates on (sometimes slightly modified) quotations and explicit mentions of names and work titles, i.e., named entities that can also be relatively easily captured by machine (cf. Molz 2020, 20 and the chapter by E. Gius in this volume). A mixed methods approach is an ideal way to correlate the references subsequently identified by humans and to detect the matches by computer linguistic tools for textual comparison. If the focus is on the detection of semantic similarities, then manual annotation must be geared towards the operationalization of the target phenomena as machine readable concepts (cf. Pichler & Reiter 2021).

2. Computational Analysis in the Reuse of Texts

A central form of digital modeling and intertextual analysis has thus far been research of text reuse: “Text reuse refers to citing, copying or alluding text excerpts from a text resource to a new context” (Moritz et al. 2016, 1849). Concrete references between texts are digitally modeled, primarily at the grammatical level, i.e., direct quotations and slightly modified paraphrases, which are possible automatically detected matches on the linguistic level of the surface of the text.[7] Various computational methods of text mining and natural language processing are used for this purpose, some of which are adopted from the field of plagiarism detection.

At the interface of literary and theological research, the DFG project *Zitieren als narrative Strategie*[8] (Citation as Narrative Strategy) at the University of Konstanz is investigating questions of cultural hybridization between classical antiquity and Christianity using a mixed methods approach for the computer aided detection of citations

6 Cf. section 3 on automated detection of semantic text similarities.

7 Büchler et al. designates these references of “paraphrasing” as “a hyponym of text reuse” (Büchler et al. 2014, 221).

8 Cf. https://www.litwiss.uni-konstanz.de/latinistik/forschung/forschungsprojekte (Accessed: 18 June 2024).

in the corpus of letters of the church father Jerome. Corpus linguistic methods such as keyword in context and part of speech analyses as well as topic modeling are used as methods that are more strongly related on the level of content (cf. Revellio 2022, 94 f.).

The methods used to analyze text reuse are primarily aimed at the automated detection of textual relationships in large text corpora. Theoretically, the concept of text reuse is based on a distant reading approach, an approach that began with the work of Franco Moretti, Matthew Jockers, and others (Moretti 2000; Jockers 2013). The automated machine detection of textual matches provides reliable mass data alongside previous individual case research, covering larger volumes of text and longer historical time spans (cf. Liebl & Burghardt 2020, 58). The Big Data argument of Moretti and Jockers is contrasted with a focus on highly canonical texts, such as the Bible and Shakespeare plays (cf. e. g. Büchler et al. 2014, Moritz et al. 2016; Liebl & Burghardt 2020). This points to a fundamental challenge of algorithm-supported research: the increased possibilities of quantitative analyses correspond to an equally increased need for example or training data, which can be used to train the algorithms and test them in terms of performance and reliability. For the phenomenon of intertextuality in the specific form of explicit references on a linguistic level, corpora exist for which references are available and known to the necessary extent, especially in the field of ancient philological and theological texts and, due to the large number of explicit references and the extensive academic debate that has been going on for centuries, in the field of Shakespeare research.

Theoretical and methodological perspectives are intertwined in the formation of the concept of text reuse: the limitations of automated evaluation necessitate a concept of intertextuality that prioritizes concrete quotations over *softer*, semantic parallels that are difficult to identify automatically (cf. Büchler et al. 2014, 221). At the same time, the use of computational methods enables one to uncover linguistic similarities that cannot be explicitly detected by human readers (cf. Coffee 2018, 207).

Some intertextuality concepts (Genette [1982] 1993; Broich & Pfister 1985; Holthuis 1993) based on precision and concretization ultimately prove to be conceptually abstracted from the focused, more linguistic understanding of correspondence on which the operationalizations under the term *text reuse* are based. Although the phenomena studied under this label can be described as the narrowest form of intertextuality, as defined by Genette in his subtypes of transtextuality (cf. Coffee 2018; Liebl & Burghardt 2020, 58), the term *intertextuality* rarely appears in studies on text reuse, despite obvious conceptual reference points (e. g., Büchler et al. 2014; Moritz et al. 2016). However, the starting point of the research are text critical questions about textual dependencies and transmission histories (cf. Moritz et al. 2016, 1894; Coffee 2018).

3. Automatic Detection of Semantic Textual Similarities

From the perspective of literary studies, which originally shaped and formulated the concept of intertextuality, computational approaches aimed at citations and linguistic similarities lack complexity (cf. Horstmann et al. 2023, 1). Computational approaches to exploring textual similarities beyond the linguistic level of grammatical parallels are emerging. These approaches run parallel to and partially overlap with the concepts of computational detection of text reuse and undertake classical intertextuality research by focusing on the analysis of content dependent on interpretation and stylistic similarities, which are central to the concepts following Genette's structuralist hermeneutic hyper textuality approach.[9] Computational analysis works seamlessly with a significant expansion of the scope of the term intertextuality more in line with Kristeva's post-structuralist/deconstructivism (cf. Scheirer et al. 2016, 205f.).

Topic Modeling is a standard method of the Digital Humanities used to detect intertextual relationships (cf. the chapter by M. Althage in this volume). In the *Intertextual Hub*[10] developed at the University of Chicago, Topic Modeling is not aimed at investigating the references and transformations of specific texts intended by authors; rather, digital intertextuality research uses the general affinity of computational methods for the comparative examination of texts to determine content-related similarities in larger text corpora. The method aligns with a stabilization of the corpus. The *Intertextual Hub* offers various collections of French literary and political texts from the 18th century, which have certain thematic and ideological similarities. The *Tesserae*[11] platform developed at the University of Buffalo focuses on ancient Latin and Greek texts as well as on texts that can be analyzed across languages regarding linguistic, semantic, metric, and phonological similarities using an online platform (cf. Coffee 2018, 207).

Alternatively, another option for the quantitative modeling of intertextual relationships is to train an algorithmic model specifically for the detection of textual similarities. The trained algorithm then has more flexible applicability, given the trained model can be applied to corpora. At the same time, however, the high degree of variability of intertextual spellings presents a challenge. Intertextuality cannot be reduced to a fixed set of text structures but can be realized at any level of the text (cf. Karrer 1985). For model training, a reduction must be made to make the textual similarities operational in the context of training data. The *CompAnno*[12] project sim-

9 Genette's theory of hypertextuality is categorized into forms that imitate the literary style of another text and those that transform the plot (cf. Genette [1982] 1993, 36–43).

10 See https://intertextual-hub.uchicago.edu (Accessed: 18 June 2024).

11 See https://tesserae.caset.buffalo.edu (Accessed: 18 June 2024). On the functionalization of *Tesserae*, see Coffee 2018, 211f.

12 Comparative Annotation to Explore and Explain Text Similarities (CompAnno); Project in the framework of the DFG focus program on Computational Literary Studies, https://dfg-spp-cls.github.io/projects_en/2020/01/24/TP-CompAnno (Accessed: 18 June 2024).

ilarly concentrates on automated detection of similarities in the representation of characters in literary texts. The intended detection and classification of similarities in this automation approach no longer refers to historically verifiable references but to model intertextuality, which is exclusively based on detectable semantic similarities.

4. Summary

Digital modeling corresponds to the general idea of the concept of intertextuality by establishing the capability of systematization of the types of writing. Conversely, the methods of Digital Humanities have an affinity to a comparative analysis of texts. A challenge remains in analog intertextuality research since there are different specifications regarding the concrete modeling of intertextual relationships, which are generally too unspecific for digital operationalization; however, the research creates a larger scope for various digital approaches.

Digital forms of differentiated manual modeling and formalization of intertextual relationships aimed at subsequent computer aided evaluation are closely linked to the literary tradition.

The reduction of the concept of intertextuality by concentrating on linguistically locatable cases such as direct quotations, paraphrases, and syntactic correspondences enables one to use a variety of established methods from computational linguistics for the automated detection of intertextuality.

The expansion of the concept of intertextuality in the direction of a general similarity of textual properties allows one to focus increasingly on the level of meaning bearing text structures within the framework of machine detection yet tends to balance out the historical dimension of textual dependencies.

Digital intertextuality research thus reproduces the inherent tension in the analog approach between the concentration on precisely determinable and marked individual text references and the perspective of a general intertextuality of literary text production. However, one advantage of digital approaches that are strongly oriented towards practical work on and with digital corpora is that different approaches to the phenomenon of intertextuality do not have to remain unconnected to each other but can be related to each other via data comparison. For example, direct quotations at a word level and detected parallels at plot or character representation level could be annotated together in a text and examined for overlaps and deviations. Conversely, automatically generated annotations can be transferred into the formalized structures of a machine-readable description system of intertextual writing styles. In this way, findings on the form and function of intertextual relationships can productively complement each other on the text level. On a metalevel, the combination of different approaches allows the associated ideas of intertextuality to be discussed. In this sense, the digital modeling of intertextuality can also contribute to the theoretical foundation of the digital practice of comparison.

References

Berndt, F., & Tonger-Erk, L. (2013). *Intertextualität. Eine Einführung.* Berlin: Erich Schmidt Verlag [= *Grundlagen der Germanistik,* 53].

Brodie, Th.L. (2006). The Triple Intertextuality of the Epistles. An Introduction. In Id., D.R. MacDonald & S.E. Porter (Eds.), *The Intertextuality of the Epistles. Exploration of Theory and Practice* (pp. 71–89). Sheffield: Sheffield Phoenix Press.

Id., MacDonald, D.R., & Porter, S.E. (2006). Introduction. Tracing the Development of the Epistles. The Potential and the Problem. In eid. (Eds.), *The Intertextuality of the Epistles. Exploration of Theory and Practice* (pp. 1–9). Sheffield: Sheffield Phoenix Press.

Eid. (2006): Conclusion. Problems of Method. Suggested Guidelines. In eid. (Eds.), *The Intertextuality of the Epistles. Exploration of Theory and Practice* (pp. 284–296). Sheffield: Sheffield Phoenix Press.

Broich, U., & Pfister, M. (Eds.). (1985). *Intertextualität. Formen, Funktionen, anglistische Fallstudien.* Berlin/Boston: De Gruyter [= *Konzepte der Sprach- und Literaturwissenschaft,* 35]. DOI: https://doi.org/10.1515/9783111712420 (Accessed: 18 June 2024).

Büchler, M., Burns, Ph.R., Müller, M., Franzini, E., & Franzini, G. (2014). Towards a Historical Text Re-use Detection. In Ch. Biemann & A. Mehler (Eds.), *Text Mining. From Ontology Learning to Automated Text Processing Applications.* FS Gerhard Heyer (pp. 221–238). Cham: Springer [= *Theory and Applications of Natural Language Processing*]. DOI: https://doi.org/10.1007/978-3-319-12655-5 (Accessed: 18 June 2024).

Coffee, N. (2018). An Agenda for the Study of Intertextuality, TAPA, 148(1), 205–223. DOI: https://doi.org/10.1353/apa.2018.0008 (Accessed: 18 June 2024).

Dörr, J.-E. (2012). *“Denn es steht geschrieben…”. Intertextuelle Bezüge im Fachtextsortennetz Andachtskalender am Beispiel des Neukirchener Kalenders.* Hamburg: Verlag Dr. Kovač [= *Angewandte Linguistik aus interdisziplinärer Sicht,* 44].

Genette, G. (1993). *Palimpseste. Die Literatur auf zweiter Stufe.* Aus dem Französischen von W. Bayer & D. Hornig. Frankfurt a.M.: Suhrkamp.

Gillmayr-Bucher, S. (2006) Intertextuality. Between Literary Theory and Text Analysis. In Th.L. Brodie, D.R. MacDonald & S.E. Porter (Eds.), *The Intertextuality of the Epistles. Exploration of Theory and Practice* (pp. 13–23). Sheffield: Sheffield Phoenix Press.

Hohl-Trillini, R., & Quassdorf, S. (2010). A ‘key to all quotations’? A corpus-based parameter model of intertextuality, *Literary and Linguistic Computing,* 25(3), 269–286. DOI: https://doi.org/10.1093/llc/fqq003 (Accessed: 18 June 2024).
Holthuis, S. (1993). *Intertextualität. Aspekte einer rezeptionsorientierten Konzeption.* Tübingen: Stauffenburg-Verlag [= *Stauffenburg Colloquium,* 28]
Horstmann, J., Lück, Ch., & Normann, I. (2023). Textliche Relationen maschinenlesbar formalisieren. Systeme der Intertextualität. In P. Trilcke, A. Busch & P. Helling (Eds.), *DHd 2023. Open Humanities Open Culture.* Trier/Luxemburg: Zenodo. DOI: https://doi.org/10.5281/zenodo.7715368 (Accessed: 18 June 2024).
Jockers, M. (2013). *Macroanalysis. Digital Methods & Literary History.* Urbana/Chicago/Springfield: University of Illinois Press.
Karrer, W. (1985). Intertextualität als Elementen- und Struktur-Reproduktion. In U. Broich & M. Pfister (Eds.), *Intertextualität. Formen, Funktionen, anglistische Fallstudien* (pp. 98–116). Berlin/Boston: De Gruyter [= *Konzepte der Sprach- und Literaturwissenschaften,* 35]. DOI: https://doi.org/10.1515/9783111712420 (Accessed: 18 June 2024).
Kristeva, J. (1972). Bachtin, das Wort, der Dialog und der Roman. Aus dem Französischen von M. Korinman & H. Stück. In J. Ihwe (Ed.), *Literaturwissenschaft und Linguistik. Ergebnisse und Perspektiven.* 3. *Zur linguistischen Basis der Literaturwissenschaft II* (pp. 345–375). Frankfurt a.M.: Athenaum.
Lachmann, R. (1990). *Gedächtnis und Literatur. Intertextualität in der Russischen Moderne.* Frankfurt a.M.: Suhrkamp.
Liebl, B., & Burghardt, M. (2020). “Shakespeare in the Vectorian Age”. An evaluation of different word embeddings and NLP parameters for the detection of Shakespeare quotes. In *Proceedings of LaTeCH-CLfL 2020* (pp. 58–68). Online: International Committee on Computational Linguistics. URL: https://aclanthology.org/2020.latechclfl-1.7 (Accessed: 18 June 2024).
Molz, J. (2020). *A Close and Distant Reading of Shakespearean Intertextuality. Towards a mixed methods approach for literary studies* [Diss.]. München: University Library of Ludwig-Maximilians-Universität München. DOI: https://doi.org/10.5282/oph.4 (Accessed: 18 June 2024).
Moretti, F. (2020). Conjectures on World Literature, *New Left Review,* 1, 54–68. URL: https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature (Accessed: 18 June 2024).
Moritz, M., Wiederhold, A., Pavlek, B., Bizzoni, Y., & Büchler, M. (2016). Non-Literal Text Reuse in Historical Texts. An Approach to Identify Reuse Transformations and its Application to Bible Reuse. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1849–1859). Austin: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/D16-1 (Accessed: 18 June 2024).
Nantke, J., & Schlupkothen, F. (2018). Zwischen Polysemie und Formalisierung. Mehrstufige Modellierung komplexer intertextueller Relationen als Annäherung

an ein literarisches ‘Semantic Web’. In G. Vogeler (Ed.), *DHd 2018. Kritik der digitalen Vernunft* (pp. 345–349). Köln: Zenodo. DOI: https://doi.org/10.5281/zenodo.3684897 (Accessed: 18 June 2024).

Ead. (2019). FormIt. Eine multimodale Arbeitsumgebung zur systematischen Erfassung literarischer Intertextualität. In P. Sahle (Ed.), *DHd 2019. multimedial & multimodal. Konferenzabstracts* (pp. 289–291). Frankfurt a. M./Mainz: Zenodo. DOI: https://doi.org/10.5281/zenodo.2596095 (Accessed: 18 June 2024).

Pichler, A., & Reiter, N. (2021). Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse. Eine Annäherung über Norbert Altenhofers hermeneutische Modellinterpretation von Kleists Das Erdbeben in Chili, *Journal for Literary Theory*, 15(1–2), 1–29. DOI: https://doi.org/10.1515/jlt-2021-2008 (Accessed: 18 June 2024).

Pfister, M. (1985). Konzepte der Intertextualität. In U. Broich & M. Pfister (Eds.), *Intertextualität. Formen, Funktionen, anglistische Fallstudien* (pp. 1–30). Berlin/Boston: De Gruyter [= *Konzepte der Sprach- und Literaturwissenschaft*, 35]. DOI: https://doi.org/10.1515/9783111712420 (Accessed: 18 June 2024).

Revellio, M. (2022). *Zitate der ‘Aeneis’ in den Briefen des Hieronymus. Eine digitale Intertextualitätsanalyse zur Untersuchung kultureller Transformationsprozesse.* Berlin/Boston: De Gruyter [= *Philologus. Suppl.*, 17]. DOI: https://doi.org/10.1515/9783110760828 (Accessed: 18 June 2024).

Riffaterre, M. (1984). Intertextual Representation. On Mimesis as Interpretive Discourse, *Critical Inquiry*, 11(1), 141–162.

Scheirer, W., Forstall, Ch., & Coffee, N. (2016). The sense of a connection. Automatic tracing of intertextuality by meaning, *Digital Scholarship in the Humanities*, 31(1), 204–217. DOI: https://doi.org/10.1093/llc/fqu058 (Accessed: 18 June 2024).

Mentioned Projects and Databases

ACDb: Augustinian Correspondence Database. Heidelberg: heiDATA, V. 3. 2016 (data set by Christopher Nunn). DOI: https://doi.org/10.11588/data/10085 (Accessed: 18 June 2024).

CompAnno: Comparative Annotation to Explore and Explain Text Similarities (Research Project by Julia Nantke and Nils Reiter from fall 2023). URL: https://dfg-spp-cls.github.io/projects_en/2020/01/24/TP-CompAnno (Accessed: 18 June 2024).

HyperHamlet. URL: http://www.hyperhamlet.unibas.ch (the page is currently under construction and is not available during this time).

TEASys: Tübingen Explanatory Annotation System. URL: http://www.annotating-literature.org/wp-content/uploads/2020/09/Styleguide-2020-08-11.pdf (Accessed: 18 June 2024).

Tesserae. URL: https://tesserae.caset.buffalo.edu (Accessed: 18 June 2024).
The Intertextual Hub. URL: https://intertextual-hub.uchicago.edu (Accessed: 18 June 2024).
WordWeb/IDEM: networks of intertextuality in early modern drama. URL: http://wordweb-idem.ch/index.html (Accessed: 18 June 2024).

Spatial Analysis, or The New Literary Geography

Matthew Wilkens

https://orcid.org/0000-0001-6749-9318

Abstract The analysis of textual geography has occupied an important place in literary and cultural studies over several decades. This chapter anatomizes the three major forms of geo-textual analysis: *thematic, deep,* and *schematic.* It finds a place for each of these forms in the history of the Digital Humanities, and argues that the schematic form of analysis is the basis of a computationally intensive New Literary Geography. Presenting case-study results from large-scale research on ethnicity and national origin in British literature, on the historical evolution of American authors' geographic attention, and on the relationship between literary and economic production, the chapter shows how a range of cultural issues can be addressed with the help of computationally produced textual-geographic evidence. It also suggests that the New Literary Geography both anticipates and precipitates concrete changes in the practice of literary scholarship – including convergence with other disciplines, increased attention to popular sources, and decreased linguistic diversity – that are now shaping the humanities as a whole.

Keywords Literary Geography, Cultural Studies, Literary Studies, Scientific Culture

Among Virginia Woolf's earliest published writing was an essay titled *Literary Geography* (1905). It was a review, in the *Times Literary Supplement,* of two biographical field guides that, she wrote, allowed readers the "sentimental" pleasure of knowing "that Thackeray rang this very doorbell or that Dickens shaved behind that identical window" (Woolf 1905, 81). She had, as we might have guessed, nothing notably kind to say about either of them.

If the work that has appeared under Woolf's heading has improved over the intervening century – and it has, to the point that it now constitutes one of the most promising avenues of socially inflected textual studies and one of the major varieties of Digital Humanities – critical reception has remained uneven. Franco Moretti's *Atlas of the European Novel* (Moretti 1998), one of the most ambitious works of literary geography to date, remains best known as a methodological precursor to the quantitative practices he would name "distant reading" two years later and demonstrate at greater length in *Graphs, Maps, Trees* (Moretti 2005). Andrew Thacker's research on the geography of modernism has been deeply influential in the relevant reaches of that subfield but hasn't traveled as widely as one might have hoped. Studies of literary

Matthew Wilkens: Spatial Analysis, or The New Literary Geography, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 295–313. DOI: https://doi.org/10.11588/heibooks.1521.c21955

regionalism of the type pioneered in the eighties and nineties by Judith Fetterley, Marjorie Pryse, and Richard Brodhead have remained primarily historical affairs. And the broad spatial turn that was an integral component of much of the theoretical energy of the last generation only very rarely became a turn to the geographic as such.

Yet literary geography is more relevant today than at any point in its long history. It is helping critics to grasp the relationship between the natural world and human society, to decipher the complex structure of textual genres and of individual narratives, to integrate readership with social and public history, to connect writing to other media arts, and to build bridges between cultural criticism and the social sciences. It is a rich set of critical practices that stretch across intellectual domains. It is also almost embarrassingly well positioned to benefit from the use of digital and computational tools, and it has, in consequence, become one of the most important areas of the Digital Humanities. If we want to know where Digital Humanities and cultural studies alike are headed in the next decade – toward closer integration with media studies and with the social sciences, toward a fuller incorporation of popular sources and genre fiction, away from multilingual comparatism – we could do worse than to understand how literary geography came to occupy this position and what its unique affordances allow, as well as what they obscure.

1. What is literary geography?

If literary geography is so important, why haven't more people heard of it? Part of the reason is that it hasn't always been clear what literary geography as a field, rather than as a series of isolated results, *is* or what one can do with it. Discounting biographical studies of the type that failed to rouse Woolf at the start of her career, geographical engagement with literature and other narrative texts generally falls into one of three categories. These could be called *thematic, deep,* and *schematic. Thematic* work is devoted to geography and space as elements of textual content, the things that readers can see and understand if they read certain books in the right way. *Deep* literary geography – or deep or thick literary mapping, as it is more commonly called – is about assembling and exploring the networked layers of cultural material that can be attached to textual places. *Schematic* analysis involves explicit modeling of texts' geographic content, usually with an eye toward scale and formal comparison. The last two of these (the deep and schematic forms), and especially the last one, are at the center of the new literary geography. It probably should also be said that, while a concern for texts unites all three areas as domains of humanistic studies, any purported line between literary geography and the discipline of geography proper will be an indistinct one.

Criticism that performs close readings of geography as theme is what we might call the old literary geography. If "old" sounds pejorative, substitute "established," or "proven," or the like; the old literary geography is old only in the sense that it has been practiced longer (by a generation or more) than the new literary geography. The idea is simply that there is established critical interest in the ways that texts and authors are shaped by place and in the ways, they shape readers' experiences of geographic space. Raymond Williams' *The Country and the City* (Williams 1973) was an important early example, tracing how his titular terms organized centuries of English literature in the service of capitalism. I have already mentioned the American reorganization of nineteenth-century local color fiction under the banner of regionalism that took place in the 1980s and '90s, led by Fetterley, Pryse, and Brodhead. The theoretical investments of traditional literary geography more broadly are many, but Mikhail Bakhtin and the idea of the chronotope (the distinctively inseparable mixture of time and space in narrative modes) loom large for many, as does Martin Heidegger's uncomfortable analysis of the links between land, place, and culture spread across many of his essays.

But it's not as though literary geography of the classic sort had a moment near the heyday of theory in the United States and then disappeared. Hsuan Hsu's *Geography and the Production of Space in Nineteenth-Century American Literature* has made a major recent intervention in the ongoing reconsideration of American literature and transnationalism, doing so by way of an explicit embrace of methods borrowed from cultural geography. Barbara Piatti has been exploring the geography of European literature for most of the last two decades. There is an active journal named *Literary Geographies.*[1] The list is an easy one to extend, but the point isn't to create a catalog. It is only to observe that a literary geography built on close readings that emphasize the role of geography as a structural and thematic element of major texts has been a fruitful and ongoing part of literary studies and allied disciplines for more than forty years.

2. Maps in depth

If thematic literary geography is akin to conventional critical reading, deep mapping more closely resembles the intellectual ambitions of the critical edition. Just as editions have been transformed by digital media and online access, the current form of the deep map owes much to the existence of digital tools. Like the old literary geography, deep mapping seeks to understand, and often to multiply, the nuances of geographical use in relatively small amounts of text. There are practical reasons why this should be so; deep mapping is difficult, time-consuming work. But, like the best

1 For the continuing relevance of the Bakhtinian chronotope in geographic analysis, see the work of the Chronotopic Cartographies project by S. Bushell et al.: www.lancaster.ac.uk/chronotopic-cartographies (Accessed: 23 June 2024).

close readings, deep mapping is more than incidentally limited in scope. Deep maps are interpretive, layered, interconnected collections of knowledge about place. The term itself is usually attributed to William Least Heat-Moon, whose book *PrairyErth – A Deep Map* presented a model study of Chase County, Kansas by way of hand-drawn maps, literary excerpts and commonplaces, historical writing, and personal narrative. Today, deep maps generally take the base form of a recognizably cartographic map, to which are attached annotations, routes, essays, photographs, spatial reprojections and transformations, audio and video clips, and so forth. Maps with some of these features are not entirely new – see, for instance, Charles Minard's 1869 cartograph of Napoleon's Russian campaign, sometimes called the "best statistical graphic ever drawn" (Tufte 2001, 40f.), or many of the nineteenth-century thematic maps collected by Susan Schulten in *Mapping the Nation* (Schulten 2012) – but deep maps have become much more accessible as they have been freed from the stasis (and the expense) of print.

Deep maps can be as simple – and as shallow – as a set of pins placed on an arbitrary base map to indicate the locations where a book's events take place. Maps of this type are common pedagogical tools, but they're also often useful to working scholars as *aide-mémoires* and as visual plot summaries (cf. the chapter by J. Peters in this volume, pp. 321–322). Deep maps can be as complex as fully modeled three-dimensional cities rendered through gaming engines and virtual reality systems (Harris et al. 2016). Most fall somewhere in between, often using *Google Maps* or *Earth* to provide annotated location information along with contemporary street view images, historical photos, and other media (see, e.g., Thomas Bruce Wheeler's *The Mapped London of Sherlock Holmes,* 2016).

Overall, however, deep maps remain, even today, more deeply theorized (David Bodenhamer and Todd Presner are major figures, Bodenhamer 2010; Presner 2014) than comprehensively executed. This is true in part because systems of academic credit do not accommodate them especially well (as is similarly true of critical editions, cf. the chapter by A. von Stockhausen in this volume, p. 341), but also because it is much easier to speculate about what deep maps *could* allow than it is to build them. Even Piatti et al.'s excellent and boundlessly inventive *Literary Atlas of Europe*[2] remains a series of suggestive sketches toward an imagined whole that has yet to emerge.[3] Products beyond *Google*'s, including *Neatline, Historypin, Peripleo,* and *ArcGIS StoryMaps,* have helped reduce some of the barriers to entry in the field, with results strongly reflected in the work of cultural institutions (where large archival holdings and a mandate for public engagement are good fits for the effort involved) and interdisciplinary grant applications (for similar reasons).

2 See http://www.literaturatlas.eu (Accessed: 23 June 2024).

3 For useful reflections on the challenges and affordances of deep literary mapping, see also Barker et al. and the essays collected in *Literary Geographies* 9.1, "Mapping as Process" (2023).

3. Geographic patterns and the new literary geography

Geographic reading and deep mapping are labor-intensive practices. They are generally useful, at least for individual researchers and critics, only when dealing with a few texts of special interest. That state of affairs describes much of what both theology and literary studies have always done, which is why literary geography of the older and deeper types has been relatively well assimilated to those disciplines over the last few decades. But related forms of spatial thinking and the same advances in computation that have made deep mapping increasingly tractable have also spurred a fundamentally different critical relationship to the geography of literature. This is what I earlier called "schematic" literary geography. It is also, increasingly, what is meant by the term computational literary geography.

Schematic literary geography is concerned with patterns of geographic attention, almost always as revealed across multiple texts. Its best-known instance is Moretti's work in *Atlas of the European Novel* (Moretti 1998) and *Graphs, Maps, Trees* (Moretti 2005). In each of those books, Moretti examined the geography of European novels, concentrating not on explications of lived spatial experience, nor on a deeper understanding of geographic settings, but instead on abstract relations between geospatial entities.

Significantly, however, Moretti's literary geography in *Atlas* and *Graphs,* while often quantitative, was never computational. His analyses were built on readings that emphasized geographic relations, readings that could presumably be carried out more quickly than most others and that, therefore, could be extended to dozens of novels in the span of a few pages. Computational text analysis, which has taken a prominent place alongside hand-extracted literary-geographic data over the last decade, is both more and less than this. Nuances of affect and irony that are clear to human readers can be difficult to detect algorithmically (cf. the chapter by R. Sprugnoli in this volume, p. 269), and even the mid-distance schematism characteristic of Moretti's work on the subject is a challenge that few computational studies have attempted. And yet computation makes truly large-scale corpora tractable. Ryan Heuser and his colleagues at Stanford have combined natural language processing with historical gazetteers to study the emotional valences of London places in nearly 5,000 eighteenth- and nineteenth-century novels (Heuser et al. 2016a; b). Ian Gregory, David Cooper, and colleagues have been engaged in a years-long series of projects to scale up their work on the literary geography of the English Lake District (Rayson et al. 2017; Cooper & Gregory 2011). Blevins (2014) used computational methods to extract place name mentions in more than 20,000 pages of nineteenth-century Texas newspapers, tracing important shifts in both regional attention and the construction of American identity over time.

Elizabeth Evans and I have used computational techniques including named entity recognition, automated geocoding, and statistical analysis to help understand the intersections of genre, ethnicity, and national origin in British fiction of the long modernist era (1880–1940). Our methods are typical of those that have dominated recent quantitative literary geography, though we have been perhaps more explicit than most in our desire to reassess existing literary and cultural claims in light of new, large-scale geographic evidence. We have shown that canonical modernism's much-noted international turn was smaller and arrived later than the same phenomenon in the full run of the period's fiction (Evans & Wilkens 2018). We also note important differences between the geographic attention of native-born white British writers and that of foreign-born authors, white and nonwhite alike. And we argue, contrary to much of the foundational thought in modernist studies, that the period is best understood in terms that deëmphasize discontinuity and rupture.

To see how computational literary geography enables such claims, as well as the challenges it involves, consider the development of our research. We began with two goals: to test the modernist internationalism hypothesis and to assess some of the differences and commonalities between native and foreign writers in the period. We assembled four sets of texts, each consisting of digitized books originally published between 1880 and 1940. These sets (called corpora) ranged in size from as many as 7,399 volumes to as few as 131. Each reflected a version of part of the literary field at the time. The largest contained all the era's British fiction held by the *HathiTrust* digital library; another was an expanded representation of authors and texts widely recognized in the existing critical literature; one comprised books by foreign writers who lived, for at least some years, in Britain; the fourth was drawn from bibliographies of London regionalist fiction.

From each of these corpora, we used natural language processing techniques to extract the names of locations mentioned in the texts, then paired those names with detailed geographic records that allowed us not only to place them on maps, but also to order them within political-administrative hierarchies (Trafalgar Square is a public space in London, England, UK). This allowed us to measure, e.g., the fraction of location mentions in each text that fell within and outside the borders of modern Great Britain. The results are shown in fig. 1.

What else does our data help us see? For one thing, that international attention was common decades earlier in non-canonical volumes (fig. 1a) than in books by well-known authors (fig. 1b). This in turn suggests that at least some important aspects of literary modernism were circulating widely in popular fiction before they appeared in the more widely studied books that critics often associate with modernism proper.

There were also important differences between different kinds of London writing. Fig. 2 shows the centers of gravity, for London locations, of the four corpora before and after 1914. The specific location of each center isn't independently important, since it represents the average of a large set of points, but the positions of the centers relative to one another capture significant aggregate differences in geographic attention between the corpora. Notable is the westward bias of geographic attention (toward wealthier areas of the city) in books by prominent authors compared to that in the larger run of British fiction and, especially, to that of London regionalist fiction.

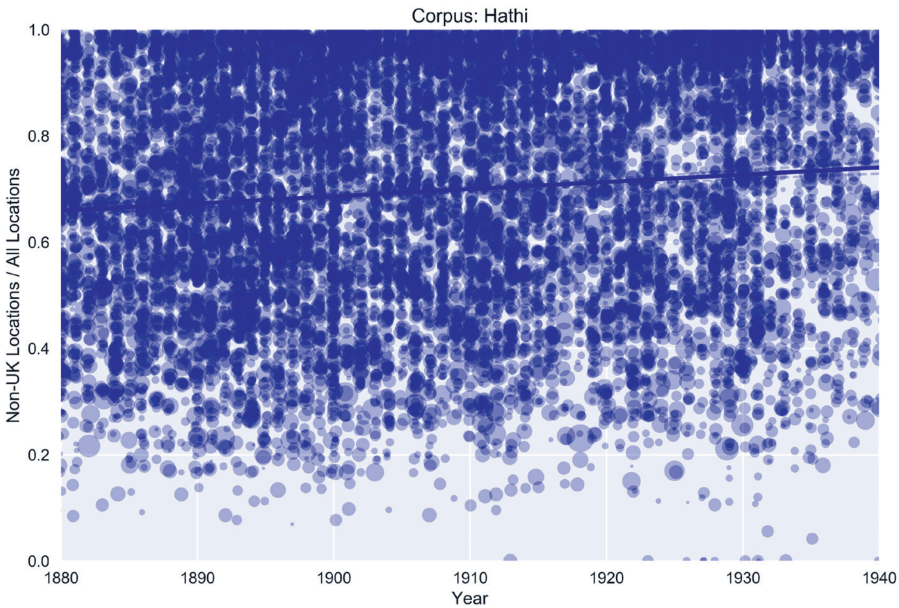


Fig. 1a Fraction of location mentions outside the United Kingdom in 7,399 volumes of British fiction, grouped by volume and ordered by publication date.

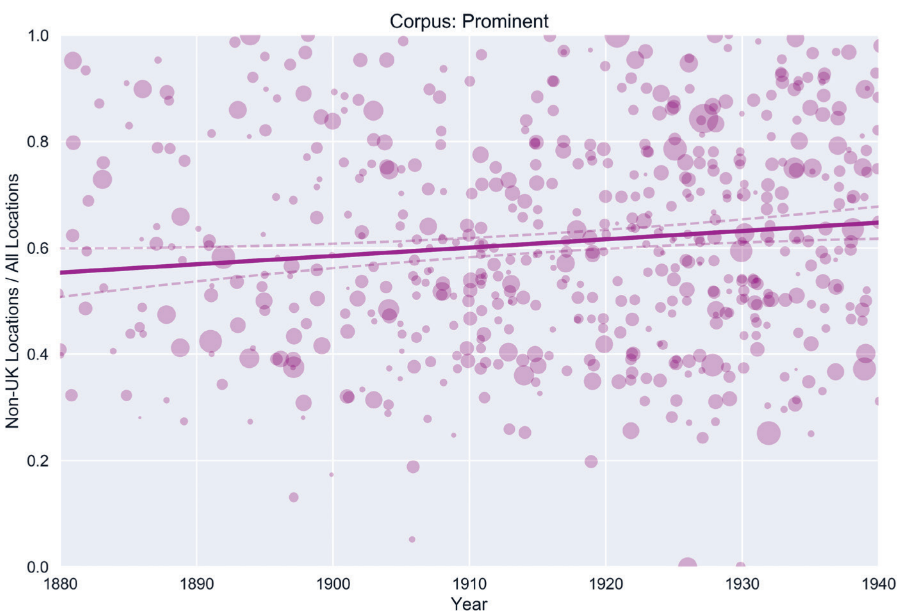


Fig. 1b Fraction of location mentions outside the United Kingdom in 576 volumes of fiction by prominent British writers, grouped by volume and ordered by publication date.

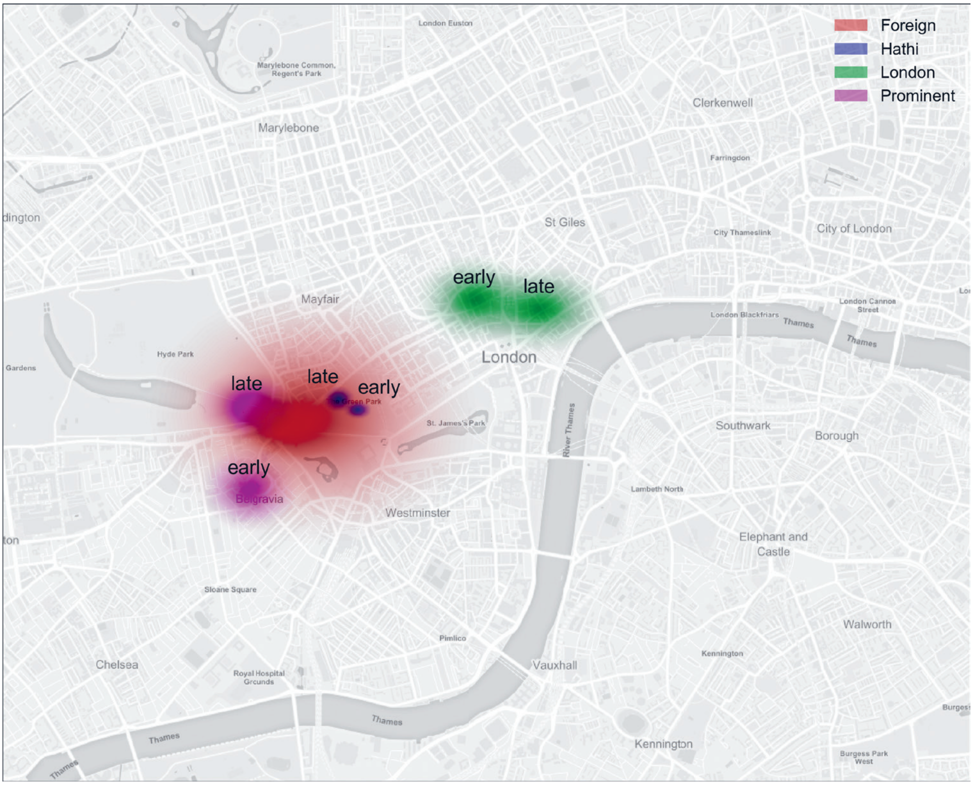


Fig. 2 Centers of gravity for London locations in the four corpora.

The regionalist corpus is distinguished in part by its higher concentration of mysteries and detective stories, as well as its greater inclusion of what we might call sociological fiction concerned with the (often sensationalized) lives of the poor. The corpus of foreign writing is the most diverse of the bunch and the only one not significantly distinct from the others.

A methodological aside: The clouds around each center represent levels of statistical uncertainty, fading to zero shading where there is less than a 1% chance of finding the true center. The most probable center is left deliberately unindicated to emphasize the probabilistic nature of these regions. It is my sense that research in the Digital Humanities is becoming more aware of the value of statistical analysis in conjunction with quantitative methods, though there is little doubt that, as a field, we are behind our friends in the social sciences in this regard.

Finally, our work on the geography of modernist-era literature serves as a reminder that, while there certainly are aspects of fiction that change in important ways over time, it can be difficult to find sharp periodizing discontinuities in large literary corpora. Part of this is a matter of recalibrating expectations. *Mrs Dalloway* is genuinely, deeply different from *Bleak House.* But it's almost impossible to imagine that modernist fiction in sum should differ from the Victorian novel to anything like the same extent. Critics know this, yet the magnitude of the difference between the cases can be difficult to appreciate until it becomes possible to examine explicitly and more nearly comprehensively some of the features of "modernist fiction" and "the Victorian novel." A central task of Digital Humanities and literary studies alike in the years ahead will be to provide a range of answers to the question "what constitutes an important shift in a range of textual properties at scale?" There won't be a single correct answer, and the answers we do produce will likely have only a small amount to do with statistical significance. But we won't know what they are until we've produced the data and, much more importantly, made the arguments.

4. Geography as symptom

As compelling as large-scale geographic data can be for what it reveals to us directly about the shape of literary-geographic attention, perhaps its greatest promise lies in its overlap with other varieties of social and cultural information. The conjunction of literary geography and geographic aspects of other fields, from economics and urban studies to sociology and history, is the core of the truly new literary geography.

To the extent that literary studies is a field invested in deciphering the relationships between aesthetic production and the cultural contexts in which that production is embedded – and it is clear that this describes a large subset of the discipline, though not all of it – quantitative literary geography offers a unique opportunity. That's because, like literary geography itself, many social scientific questions have obvious spatial components. Urbanization, the great demographic change of the nineteenth century, can be measured and tracked via census data. Economic regionalism and globalization are the subjects of huge amounts of data-driven work in economics and history. Ethnicity, immigration, and national origin are measured and tracked by government offices, sociologists, and many scholars within the humanities proper. All these social phenomena bear on and are affected by literature. If we can assess how they co-vary, we will have produced new evidence through which to interpret the larger place of literature.

What does this work look like? Consider the problem of lag between cultural changes and their representation in literature. This lag might, in principle, run in either direction. What did it actually do? Do changes in textual attention lead or follow changes in demography? If we examine US fiction in the decades around the Civil War, we can measure how thousands of books distributed their attention to the nation's rapidly evolving cities. We can then compare the fraction of all literary location mentions devoted to a fixed set of cities (and to locations within those cities) against the fraction of US population for which those same cities accounted at different times.

Fig. 3 How well does the distribution of literary attention in a corpus of mid-nineteenth-century US novels match the population of selected US cities across many decades of census data?

So, for instance, if Chicago contained 0.8% of the US population in 1870 and accounted for 0.7% of literary-geographic attention in the Civil War-era fiction corpus, we would say that Chicago was (slightly) underrepresented in the literature. We could then extend the same analysis to more cities and more dates, seeking to identify the historical point at which literary attention most closely matched an existing urban population distribution. The result of that analysis is shown in fig. 3.

What we find is that for American novels published between 1851 and 1875, their distribution of attention to urban areas matched most closely the population of those areas as it stood in the 1830 census, falling off smoothly as one moves forward or backward in time from that date. The average publication date of a book in the corpus was 1863, implying a roughly thirty-year lag between spatial population shifts and their reflection in fiction, with the caveat that the process is not a matter of recognition switching on and off, but of gradual convergence across many hundreds of authors and books.

This provides us, in turn, with a new interpretive opportunity. How should we explain the direction and length of this lag? If we think that a writer's experiences are relevant, we might note that the average age of the authors in our corpus at the time each of their books was published was 42 years old. That would pin the lag to the school days of the average author, suggesting a role for the geographic education that was increasingly popular in the early nineteenth-century US (see Schulten 2012) and suggest a kind of post-schooling stasis in which aspects of worldview formed at a young age remain relatively stable thereafter. If we are (perhaps properly) wary of such psychologizing and drawn instead to historical explanations, we might emphasize that 1830 represents about the latest date before large-scale immigration transformed Eastern cities and that it fell at the leading edge of the rapid, sustained westward expansion that continued through the end of the century. Or readers' tastes may have been important, driving literature toward an imagined urban geography that was conservatively familiar without the estrangement of explicitly historical writing that, in the American case at the time, would have been primarily rural.

The data of literary geography do not answer these interpretive questions. They are not meant to. But they provide the impetus to see these questions as important, unresolved aspects of mid-century American literary history. If scholars are sometimes inclined to treat data-intensive inquiry as "mere" description or hypothesis testing, antithetical to the open-ended aims of the humanities (whether or not our aims are really so open-ended is a separate question; I have my doubts), here is an instance – like every other one I know in the best of the Digital Humanities – in which quantitative analysis feeds qualitative, interpretive inquiry in the most direct way.

So computational literary geography, in combination with demographic data, can help us understand the temporal dynamics of cultural uptake in fiction. What other social data might be similarly useful? The possibilities are many. We could use household income and manufacturing data to help characterize the environments in which novels are set, helping us to examine class dynamics in large corpora across long time spans. We could track changes in transportation networks and in mass media circulation to compare their effects on the experiential geography and social structure of fiction. We could look for natural experiments in market incentives or corporate structure – the introduction of new prizes or subsidies for writers, the advent of corporate conglomeration in the publishing industry (Sinykin 2023) – to see how they affect both the locations of literary production and the geospatial content of that production. Or we could use global economic and literary geographic data to assess and to reevaluate, if necessary, long-standing assumptions about the relationship between economic neoliberalism and the outlook of twentieth-century and contemporary fiction.

The last possibility, combining historical economic data with literary-geographic information to assess the impact of rising neoliberal orders on American novels, is work that now exists (Wilkens 2016a). The underlying hypothesis is that the market came to function, over the course of the twentieth century, as a horizon of possibility for thought, for experience, and for action. To the extent that this hypothesis holds at a given moment, we would expect aspects of literature to resemble aspects of markets. Geography gives us a useful way to quantify the nature and extent of one version of this resemblance. As the distribution of economic output shifted markedly in the twentieth-century, we would expect to see the distribution of literary-geographic attention move in at least roughly similar ways and, importantly, for the two to track one another more closely toward the end of the period than they did at the beginning, as neoliberalism became ever more hegemonic. What we observe, using a corpus of

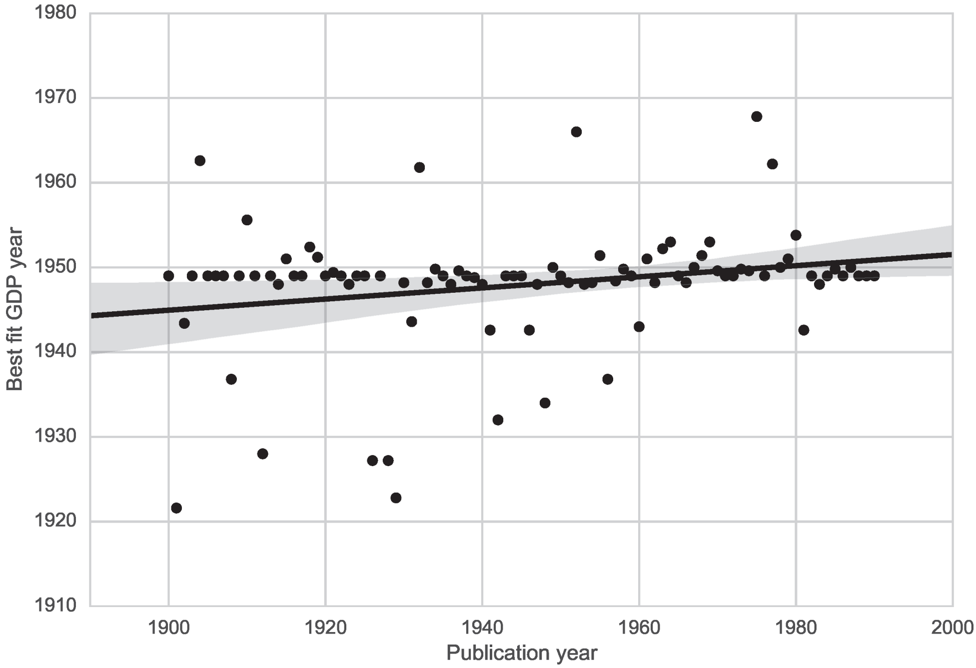


Fig. 4 Average year of closest fit between global GDP distribution and global literary-geographic attention as a function of literary year of publication.

6,942 US novels published between 1900 and 1990 and the historical GDP shares of ten major nations, is that, as far as American authors were concerned, it was always 1950 (see fig. 4). A large sample of US fiction, published both before and after 1950, distributed its global geographic attention in ways that most closely matched the distribution of global economic output as it existed immediately after the Second World War. It's impossible to assess whether changes in economic output became more or less closely aligned with changes in collective literary output over time because, when it came to international literary attention among US authors, there was no real change over time.

What's going on here? The issue, in brief, is that American authors wrote overwhelmingly about locations in the United States, consistently devoting about 80 % of their place name mentions to domestic toponyms. Economic geography shows far more variability, but that variation simply isn't reflected in American literary-geographic practice. The years between 1945 and 1950 just happen to have been the peak of US global GDP share (on the order of 40 % in 1945), meaning that those were the years when the two distributions were most closely aligned.

This finding of literary-geographic stasis is important for at least three reasons. For one, it reveals a large difference between domestic and international responsiveness in aspects of US fiction. That is, while it is relatively easy to find significant redistributions of attention within the United States in response to domestic demographic and economic changes, American novels appear to have maintained across almost the whole of the twentieth century a much more static (and generally inward-looking) treatment of the world (see also Wilkens 2021). Second, this situation produced in effect a growing conservatism on the part of American literature in the postwar era as those books' perpetual disregard for the world beyond US borders grew increasingly out of step with the global system of economic production. Finally, we must concede that, in this case, it is difficult to find support for a critique of neoliberalism that understands its object to take the form of an inescapable horizon of thought or condition of possibility. This isn't to say, wildly implausibly, that neoliberalism doesn't exist. Nor does it constitute some deep blow to its ontological interpretation. But it does provide a counterexample to think with, a case in which we were able to make a clear prediction following from a theoretical position yet failed to observe the predicted result. I'm a Bayesian, not a Popperian; there's no falsification here, but I think we should be willing to update our critical priors.

5 New literary geography, new literary studies

Where does all of this leave us as we contemplate the near future of the Digital Humanities and of humanistic disciplines alike? As I suggested at the outset, I see three clear implications from the rise of the new literary geography.

5.1 Convergence with adjacent fields

Literary geography is attractive to many scholars because the geography of literature is interesting and important in its own right. It allows critics to provide better answers to complicated interpretive questions and to track underlying contrasts in the ways different literary traditions relate to the environments around them. This is work that can often benefit from the addition of computational techniques, though it is no more constitutively computational than most other areas of literary studies. The advantages of computation in both cases – for literary geography and for literary studies in general – are, at first order, those of scale (cf. also the chapter by W. Mattingly in this volume, pp. 177–179). This is a long-standing point, to which one need only add two small elaborations. One, scale is a relative term. There exists excellent, persuasive, computationally assisted work on corpora as small as a handful of novels or the set of Shakespeare's plays. In literary geography, Cooper & Gregory's (2011) studies of the English Lake District bring new insights using as few as two book-length texts. Small-corpus work highlights the fundamental interplay of scales that is almost always at work in literary scholarship and that continues to inform quantitative studies. And two, even when the computational scales involved are much larger, critics will often return to individual authors and texts to read them in light of their place in a newly visible context. Virginia Woolf's geographic attention, for example, was very different from that of most of her peers, canonical and otherwise (more specific and London-centric within Britain, less international and less geographically intensive overall; see Evans & Wilkens 2018). The outline of her differences isn't altogether surprising, though we should be wary of our ability to offer easy post hoc claims about whatever we observe. But the magnitude of her difference is genuinely unexpected; quantification helps us to recognize that Woolf's work represents a species of geographic pole within modernist literature and to revisit it and its interlocutors accordingly.

The core of the new literary geography, however, is only partly about deeper access to individual texts or to geography as such. It is also, even primarily, about the production of spatialized data in support of critical cultural analysis. By that I mean that the new literary geography represents an extension of the cultural turn that has driven much of the best, most important literary scholarship of the last two generations. Literary texts remain at its heart because literature is a notable site of cultural production, one with its own history, its own forms, and its own interpretive tradition. But the largest goals of the new literary geography have much more to do with describing and explaining cultural formations than they do with any final investment in literature per se, even when they proceed entirely by way of literary texts.

5.2 A larger role for popular and genre literature

While the advantages of computation aren't solely those of scale, the fact that computers can deal with billions of words interacts constructively with the broader cultural turn just noted. As literary scholars have come increasingly to treat books as socially symptomatic, the logic of limiting critical work to dozens or hundreds of widely read texts has been progressively eroded. There is no reason, in principle, to believe that an aesthetically good book will tell us more about the culture that produced it than will a bad one. Computation, in turn, has chipped away at the remaining practical justifications for strictly canonical work. As we have seen in the examples of what I've called schematic literary geography, there is much to be gained through the relatively new ability to situate known authors and texts within a larger literary field that includes far more unknown and understudied work.

The larger point is that the types of computational analysis native to the new literary geography in particular and to (part of) the Digital Humanities in general have accelerated a shift already long underway toward greater engagement with what Merve Emre has recently called (in a different context) "paraliterary" reading (Emre 2017). When combined with an increased, field-wide emphasis on social and cultural questions, the future of literary studies again looks convergent with that of disciplines that share an interest in cultural production across media and from both sides of the production-reception equation.

5.3 Decreasing linguistic diversity

I take both of the preceding developments to be straightforwardly positive. They make literary studies broader, more relevant both inside and outside the university, and more interesting to more people and for more purposes. But the long-term evolution that has been assisted by computation has some obvious downsides (see also Chun et al. 2016). Among these is the lamentable waning of multilingual and comparatist work that is reflected indirectly in declining non-English language course enrollments at American universities since 1960 (over 50% per capita; Looney & Lusin 2018, 12). We should be careful about cherry-picking start dates for these sorts of comparisons, and we should acknowledge that universities are very different places, enrolling different students and offering different opportunities, especially for women and minorities, than they were in 1960. Most of the decline in foreign-language enrollment happened in the 1970s; they've remained roughly flat for the last 40 years. But let us stipulate that English hasn't exactly *declined* in prominence as a global language or within the academic modern languages of late. Most of this is down to factors far broader than the rise of what remains, today, a small subfield of a few humanities disciplines. But it is true that some techniques in computational text analysis depend on labor-intensive software development and linguistic training data that are much more likely to cover English-language texts than those in other languages. To the extent that these techniques become more important to the field, they will exert further pressure away from multilingual literary research.

That said, there are reasons for optimism on this point. For one, many computational methods are largely language-agnostic. Apart from a bit of hand-waving about tokenization, computers are happy to count words in any language and many computational techniques are, at bottom, matters of manipulating word counts. Even where language-specific training data is necessary, as it generally is in computational literary geography, there are good, open-source language models for many languages beyond English. This is especially true for the major European languages and for Chinese, Japanese, and Arabic, each of which has either (or both) a major funding constituency or a robust active user base. There will be more such resources in the future as they become more important for scholarship and commerce alike. Still, there is a real danger that languages that are currently marginal in the academy will fall further behind in the face of new barriers to entry for scholarship.

Linguistic uniformity is an area of concern for DH and for literary studies in general, both of which need urgently to become more diverse in any number of ways. We should acknowledge, though, that doing more computation may help language departments on this score more than it hurts them in the long run. It doesn't undermine near-term efforts to increase diversity to note that a more diverse DH ultimately depends on a more diverse humanities. There is evidence that majors that are perceived to lead to more immediately lucrative careers are more attractive to students of comparatively lower socioeconomic standing (Ma 2009; Morgan et al. 2013; Pinsker 2015). If the ability to integrate close reading with statistical literacy, programming, and data science makes humanities disciplines an easier sell to students from families outside the top income brackets, this is a strong mark in its favor. It is also how fields grow and adapt over time, by recruiting and retaining new people and new perspectives. This isn't the whole solution to the lack of diversity in DH and in many humanities departments, but is, I think, an important component of building the kind of vibrant community we rightly desire.

6 Conclusion

The new literary geography is, finally, an orientation toward narrative texts. It treats texts as complex, socially symptomatic objects, valuable both singularly and in aggregate, from which can be extracted information that advances the ends of literary studies as the field is configured today and in dialogue with its existing methods. While it legitimately encompasses – via deep, thick, and digital mapping – the kind of close attention to individual texts that has long characterized humanities disciplines, its most innovative aspect is its explicit embrace of new opportunities for integration with the social sciences. It is also one of the most actively evolving and expanding varieties of Digital Humanities. For both of these reasons, the new literary geography itself serves as a symptomatic proxy for the future directions of the larger literary field in which it is embedded. It portends a more vibrant and diverse humanities, though only if both areas can resist the gatekeeping urge to maintain purity under the guise of either rigor or resistance.

Thirty years ago, Fredric Jameson wrote, in a discussion of postmodernism and late capitalism, that "the decision as to whether one faces a break or a continuity […] is not an empirically justifiable or answerable one, since it is itself the inaugural narrative act that grounds the perception and integration of the events to be narrated" (Jameson 1991, xii). My sense is that discussions of the place of the Digital Humanities within the larger humanities, both pro and con, too often make that decision in favor of rupture. This is perhaps ironic in an essay devoted to defining something called the new literary geography. But as I have tried to show, computational methods in general and literary geography in particular differ substantially in detail but much less in aims and outlook from older forms of textual analysis and cultural studies. The more deeply we can appreciate the fundamental continuity between quantitative and traditional methods, the faster and better will be the evolution of textual studies as a whole.

References

Bakhtin, M. (1981). *The Dialogic Imagination. Four Essays* (C. Emerson & M. Holquist, Trans.). Austin: University of Texas Press.

Barker, E., Bouzarovski, S., Pelling, Ch., & Isaksem, L. (Eds.). (2016). *New Worlds Out of Old Texts. Revisiting Ancient Space and Place.* Oxford: Oxford University Press.

Beaumont, A., Hones, Sh., & Ridanpäa, J. (Eds.). (2023). *Literary Geographies,* 9(1). URL: https://www.literarygeographies.net/index.php/LitGeogs/issue/view/18 (Accessed: 23 June 2024).

Blevins, C. (2014). Space, Nation, and the Triumph of Region. A View of the World from Houston, *Journal of American History,* 101(1), 122–47. URL: https://www.jstor.org/stable/44285951 (Accessed: 23 June 2024).

Bodenhamer, D.J., Corrigan, J., & Harris, T.M. (Eds.). (2015). *Deep Maps and Spatial Narratives.* Bloomington: Indiana University Press.

Eid. (Eds.). (2010). *The Spatial Humanities. GIS and the Future of Humanities Scholarship.* Bloomington: Indiana University Press.

Brodhead, R. (1993). *Cultures of Letters. Scenes of Reading and Writing in Nineteenth-Century America.* Chicago: University of Chicago Press.

Chun, W.H.K., Grusin, R., Jagoda, P., & Raley, R. (2016). The Dark Side of the Digital Humanities. In M.K. Gold & L.F. Klein (Eds.), *Debates in the Digital Humanities* (pp. 493–509). Minneapolis: University of Minnesota Press. DOI: https://doi.org/10.5749/j.ctt1cn6thb.41 (Accessed: 23 June 2024).

Cooper, D., & Gregory, I.N. (2011). Mapping the English Lake District. A Literary GIS, *Transactions of the Institute of British Geographers,* 36(1), 89–108.

Emre, M. (2017). *Paraliterary. The Making of Bad Readers in Postwar America.* Chicago: University of Chicago Press.

Evans, E.F., & Wilkens, M. (2018). Nation, Ethnicity, and the Geography of British Fiction, 1880–1940, *CA. Journal of Cultural Analytics,* 3(2), 1–48. DOI: https://doi.org/10.22148/16.024.

Fetterley, J., & Pryse, M. (2003). *Writing out of Place. Regionalism, Women, and American Literary Culture.* Champaign: University of Illinois Press.

Harris, T.M., LaFone, H.F., & Bonenberger, D. (2016). From Mapping Text in Space to Experiencing Text in Place. Exploring Literary Virtual Geographies. In D. Cooper, Ch. Donaldson & P. Murrieta-Flores (Eds.), *Literary Mapping in the Digital Age* (pp. 221–239). London: Routledge.

Heuser, R., Algee-Hewitt, M., Lockhart, A., Steiner, E., & Tran, V. (2016a). Mapping the Emotions of London in Fiction, 1700–1900. A Crowdsourcing Experiment. In D. Cooper, Ch. Donaldson & P. Murrieta-Flores (Eds.), *Literary Mapping in the Digital Age* (pp. 25–46). London: Routledge.

Heuser, R., Moretti, F., & Steiner, E. (2016b). *The Emotions of London, Literary Lab. Pamphlet,* 13, 1–9. URL: https://litlab.stanford.edu/assets/pdf/LiteraryLab Pamphlet13.pdf (Accessed: 23 June 2024).
Hsu, H.L. (2010). *Geography and the Production of Space in Nineteenth-Century American Literature.* Cambridge: Cambridge University Press.
Huehls, M., & Greenwald-Smith, R. (Eds). (2017). *Neoliberalism and Contemporary Literary Culture.* Baltimore: Johns Hopkins University Press.
Jameson, F. (1991). *Postmodernism, or The Cultural Logic of Late Capitalism.* Durham: Duke University Press.
Least Heat-Moon, W. (1991). *PrairyErth. A Deep Map.* Boston: Houghton Mifflin.
Looney, D., & Lusin, N. (2018). Enrollments in Languages Other Than English in United States Institutions of Higher Education, Summer 2016 and Fall 2016. Preliminary Report. New York: Modern Language Association.
Ma, Y. (2009). Family Socioeconomic Status, Parental Involvement, and College Major Choices. Gender, Race/Ethnic, and Nativity Patterns, *Sociological Perspectives,* 52(2), 211–234.
Moretti, F. (1998). *Atlas of the European Novel.* London: Verso.
Id. (2005). *Graphs, Maps, Trees. Abstract Models for a Literary History.* London: Verso.
Morgan, S.L., Gelbgiser, D., & Weeden, K.A. (2013). Feeding the Pipeline. Gender, Occupational Plans, and College Major Selection, *Social Science Research,* 42(4), 989–1005.
Piatti, B. (2008). *Die Geographie der Literatur. Schauplätze, Handlungsräume, Raumphantasien.* Göttingen: Wallstein Verlag.
Pinsker, J. (2015). Rich Kids Study English, *The Atlantic.* URL: https://www.theatlantic.com/business/archive/2015/07/college-major-rich-families-liberal-arts/397439 (Accessed: 23 June 2024).
Presner, T., Shepard, D., & Kawano, Y. (2014). *HyperCities. Thick Mapping in the Digital Humanities.* Harvard: Harvard University Press. URL: https://www.hypercities.com (Accessed: 23 June 2024).
Rayson, P., Reinhold, A., Butler, J., Donaldson, Ch., Gregory, I., & Taylor, J. (2017). A Deeply Annotated Testbed for Geographical Text Analysis. The Corpus of Lake District Writing. In *Proceedings of 1st ACM SIGSPATIAL Workshop on Geospatial Humanities* (pp. 9–15). Redondo Beach: Association of Computer Machinery. DOI: https://doi.org/10.1145/3149858.3149865 (Accessed: 23 June 2024).
Schulten, S. (2012). *Mapping the Nation. History and Cartography in Nineteenth-Century America.* Chicago: University of Chicago Press.
Sinykin, D. (2023). *Big Fiction. How Conglomeration Changed the Publishing Industry and American Literature.* New York City: Columbia University Press.
Thacker, A. (2005). The Idea of a Critical Literary Geography, *New Formations,* 57, 56–73.
Tufte, E. (2001). *The Visual Display of Quantitative Information.* Cheshire: Graphics Press.

Wheeler, Th. B. (2016). *The Mapped London of Sherlock Holmes.* London Secrets. URL: https://sherlock-london.com (Accessed: 23 June 2024).
Wilkens, M. (2013). The Geographic Imagination of Civil War-Era American Fiction, *American Literary History,* 25(4), 803–840. URL: https://www.jstor.org/stable/43817603 (Accessed: 23 June 2024).
Id. (2016a). The Perpetual Fifties of American Fiction. In M. Huehls & R. Greenwald-Smith (Eds.), *Neoliberalism and Contemporary Literary Culture* (pp. 181–202). Baltimore: Johns Hopkins University Press. URL: https://post45.org/wp-content/uploads/2015/10/Wilkens.pdf (Accessed: 23 June 2024).
Id. (2016b). Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction, *Journal of Cultural Analytics,* 2(2), 1–24. DOI: https://doi.org/10.22148/16.009 (Accessed: 23 June 2024).
Id. (2021). Too isolated, too insular. American Literature and the World, *Journal of Cultural Analytics,* 6, 52–84. DOI: https://doi.org/10.22148/001c.25273 (Accessed: uune 2024).
Williams, R. (1973). *The Country and the City.* Oxford: Oxford University Press.
Woolf, V. (1905). Literary Geography, *Times Literary Supplement,* 81.

Figure Credits

All illustrations were generated by the author himself. The following list names the publications where they already appeared prior:

Fig. 1a: Evans & Wilkens 2018, 21.

Fig. 1b: Ibid., 22.

Fig. 2: Ibid., 34.

Fig. 3: https://mattwilkens.com/2015/01/13/literary-attention-lag [Blog] (Accessed: 24 June 2024).

Fig. 4: Wilkens 2016a, 195.

Computational Theology and Information Visualization

Janelle Peters

Abstract Information visualization assists in interpreting theological texts by allowing for close and distant readings. For close reading, visualization allows the comparison of manuscripts and linguistic features of a theological text. Distant readings using information visualizations of theological texts follow the schema of Moretti's graphs, maps, and trees. Oxford's *Paul and Patristics* project, helmed by Jenn Strawbridge, and *Coptic Scriptorium,* created by Carrie Schroeder, use graphs to illustrate word counts. Maps can be made using Stanford's ORBIS, an interactive map of geographical, economic, and social data of the late antique Mediterranean. Stanford's Palladio, which was originally designed to analyze social networks of elite medieval families, can visualize the location of Dead Sea Scrolls among the various Qumran caves and the Cairo Genizah. Trees have followed in the formidable legacy of Busa's *Index Thomisticus.* Advantages of these readings include the ability to assemble data on word count, variants, emendations, morphology, syntax, and the cultural backgrounds of manuscripts. Disadvantages to information visualization range from distorting the certitude of the translation or network via aesthetics to the durability of the analytical programs and final produced media. Possible future applications of information visualization in theological studies could look to the Map of Science, visualizing the networks of various researchers and their projects, and recreations of locations described by texts and locations in which texts were performed, read, and reproduced through visualization projects using technology such as augmented and virtual reality.

Keywords History of Visualization, Graphs, Maps, Trees, Close Reading, Distant Reading, Theological Data Visualization

Different forms of information visualization have been created independently of innovations in publishing (e.g., scroll, codex, printing press). From the earliest cave art onward, maps have hinted at human interactions with their physical and social contexts. Often encapsulating economic and logistical data, charts have connected different categories of data with one another, particularly after the line, bar, and pie charts of William Playfair (1759–1823).

Vast amounts of data can form an assemblage that invites the specialist or casual reader to play with different interpretations. Relational graphics such as the scatterplot, for instance, ask for interpretation about the causal relationship between X and Y (Tukey 1983, 47). The boxplot, whether as imagined by John Tukey (1915–2000) or Edward Tufte (b. 1942), allows the variation of a measurement to be seen.

At the same time, many underlying features of information visualization have been present in theological studies since quite early. Concordances have visually

Janelle Peters: Computational Theology and Information Visualization, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 315–331. DOI: https://doi.org/10.11588/heibooks.1521.c21956

tabulated the intensity of word references in theological texts since the medieval period, allowing users to interactively look up instances for themselves. Hugh of Saint-Cher (1200–1263), a French Dominican friar who would be made the first Dominican cardinal, is credited with creating the first concordance of the Vulgate and thus the Bible. He was said to have been helped by as many as 500 Dominican associates to produce the concordance in 1230, with divisions by book and chapter. Hugh was also a theological commentator, glossing and adding longer commentary to established theologians such as Peter Lombard. It is not a stretch to speculate that were Hugh to be participating in the current digital turn of theology he would be enthused about the many of the current projects connecting texts such as those of Alexander of Hales, Aquinas, biblical authors, and ancient philosophers. Rabbi Isaac Nathan created a concordance for the Hebrew Bible between 1438 and 1448 that was printed in Venice in 1524 and appeared in a Latin edition of 1556. From the sixteenth to eighteenth centuries, concordances also were important for studies in the Greek Septuagint, the Greek New Testament, and the Syriac.

Given the importance of concordances for theological studies, it is perhaps no surprise that the birth of the Digital Humanities is widely attributed to Italian Jesuit priest Roberto Busa (1913–2011), who collaborated with IBM to produce a digital concordance of the *opera omnia* of Thomas Aquinas. Busa announced in the July 1950 *Speculum* his intention to use "mechanical devices" to produce an *Index Thomisticus* (Busa 1950, 424 f.).[1] The very next year, 1951, James Collins wrote a *Speculum* review for a work by Busa on Aquinas by noting the general problem of situating Aquinas' words in an individual work within his larger life and its influences: "A comparative linguistic approach to the authors consulted by Aquinas cannot be carried out at present on any large scale, and the method cannot be applied even to the entire Thomistic corpus until such a project as Busa's Index is completed" (Collins 1951, 495 f.). Within the decade, Roy Wisbey had created a High German word index (Jones & Yeandle 2021). Within three decades, Busa, IBM, and an Italian committee of scholars had produced 49 volumes of the *Index Thomisticus,* which analyzed 179 texts (118 of the Thomistic corpus and 61 of other authors from the ninth to sixteenth centuries). In addition, there are titles, quotations, paraphrases of other authors, references, and transcriptions of the author's spoken words (Burton 1984). The legacy of this work is available on paper, on CD-ROM, and online.[2] Other projects have developed based on the work of Busa's team. One of these is the *Index Thomisticus Treebank* at the CIRCSE research center of the Università Cattolica del Sacro Cuore in Milan, Italy,[3] which analyzes Aquinas' morphology, form, and meaning (Passarotti 2019). The collaborative

1 Busa had been inspired by his experience as a doctoral student of having to produce 10,000 cards to discuss *praesens* and *praesentia* in Aquinas and how they occur with the word "in". Cf. Busa 1980, 83.

2 S. http://www.corpusthomisticum.org, accessed 24 June 2024.

3 S. http://centridiricerca.unicatt.it/circse, accessed 24 June 2024.

and yet still prestigious ethos of Aquinas information visualization is a real strength and, hopefully, a model for future scholarship.

Information visualization has advantages in close, mid-range, and distant reading of theological texts. For close reading, morphology, textual structures, and semantic and cultural meanings can allow readers to annotate the text with richer interpretations. Variant readings can be inserted into the text, and intralinear comparison to major intertext of the text might be made (cf. J. Nantke's chapter in this volume). For distant reading, it allows viewers to interrogate correspondences among data and to intellectually apprehend large quantities of data within a data set, which can bring together traditionally disparate items. Named entities of a text can be made apparent, even so-called "unnamed" figures such as the biblical "Lot's wife," "Jepthah's daughter," "Chloe's people," "Pilate's wife," or "Rufus' mother" (cf. E. Gius's chapter in this volume). Timelines have the capacity to illumine when different named entries appear in the text, allowing the reader with knowledge of the plot or document outline to better analyze when references occur.

Disadvantages of information visualization include: (1) the impression that the data and its metadata underlying the visualization are more securely labeled than they are in reality (e.g., literary allusions that could be alternatively religious founder sayings or later follower expansions are pinned down historically leading to a falsely strong picture of either the expressions of the founder or the followers in the allusion aggregation); (2) the impression that networks are more cohesive or relational than they are (e.g., a prosopography of indifferent acquaintances); (3) the impression that the categories of the data visualization are static over time and cultures (e.g., a world map of the phoenix); (4) the impression that the data visualization's categories constitute a particularly salient definition of the phenomenon (e.g., portable altars); (5) the impression that the data visualization's neat and even "beautiful" ordering makes just-so sense of tragedies and atrocities; (6) the impression that statistical frequency is more important than momentous silences or rare occurrences.

In this article, I will analyze several aspects of data visualization in theological and religious studies. First, I will describe current data visualization projects on different aspects of theology, focusing on the use of *Close Reading, Distant Reading,* and *Mapping* to make sense of large quantities of data for both scholars and the general public. Then, I will analyze some of the problems with data visualization as they occur in theological data visualization. Lastly, I will compare the work on data visualization being performed in theological and religious studies with the work on data visualization being performed in other fields (e.g., various sciences) to suggest possible avenues for development of theological data visualization such as virtual and augmented reality worship setting visualizations.

1. Information visualization with close and distant reading

1.1 Close reading visualization

Close readings exist at the story-level of the text, balancing the literal and the allegorical meanings of the text (Szondi 1978). The word's definition, its meaning in that text, and its appearances elsewhere in contemporaneous other literature (e.g., other instances of *pneuma* in Greek sources contemporary to the Gospel of Mark) – rather than contemporaneous manuscripts of the same document (e.g., 11th-century copies of the Gospel of Mark) – all matter. If objects described in the text have a somewhat standardized appearance in material culture, it is helpful to include them, though some might have wide variation (e.g., the plethora of animals on rhytons) and some might only be attested by much later artistic representations (e.g., the Ark of the Covenant).

Digital visualizations have allowed readers to click on words of the text for expanded information about meaning, grammar, statistical frequency, and other relevant information. In this capacity, data visualizations serve not only to further the goals of the exegesis of literary texts but to illuminate for individual interpreters without formal theological training what comprises critical theological interpretation. Of note is the fact that the components of theological interpretive practice are most closely related to the close reading practices of literary studies with linguistic and historical additions. When a viewer encounters a data visualization that gives a manuscript's text along with information about each lexical unit on the page, practices like sifting through the *peshat* interpretations of others along with the *derash* interpretations are thus not included.[4] These must be left for other visualizations, such as connecting the thought of contemporary authors through their shared themes and literary allusions in close, distant, or even midrange readings.[5]

Linguistic analysis is essential to close reading, and many projects allow users to explore the linguistic features of words. A particularly elegant reader is available from the Quranic Arabic Corpus[6], which is housed at the University of Leeds. The grammar, syntax, and morphology for each Quranic verse are displayed. The Quranic Arabic Corpus features the Arabic text along with the transliteration and definition, meaning that a user with no Arabic knowledge may click on the transliterated word and be provided with a list of other occurrences of that defined word. A variety of colors help to introduce relevant lexical units of information for non-specialists. As it

4 For an explanation of how the work of finding the peshat interpretations must be ongoing at the same time that previous peshat interpretations may be appreciated, see Simon & Greenstein 1988.

5 For a description of how human-edited outlines of a text might be analyzed for digital visualization, see Booth 2017.

6 See https://corpus.quran.com (Accessed: 24 June 2024).

does with classical Latin and Greek texts, *Perseus*[7] provides the standard Greek text of New Testament books and allows the user to click on each word for morphological analysis. This feature is available with paid applications for electronic readers such as the Kindle that allow the user to look up the form of any ancient Greek word in any ancient Greek book loaded onto the device, as happened on the first day of my *1 Clement* reading group. It is also available with paid products with carefully curated Greek texts such as *Logos,* which has supplanted the now-defunct *BibleWorks.* These projects are oriented toward allowing both specialists and those with little ancient language knowledge to use the induction method of engaging with the ancient language of the text. They greatly improve the speed at which someone may read any Greek text, though the semantic range of the provided definitions and the relevance of the comparanda of each word might still be improved with advanced in visual representation capacities. Advanced textual tools are available to assist specialists with analyzing user-provided segments of texts. These include those of the *Coptic Scriptorium*[8]. All of these tools allow readers to engage texts with more fluidity and knowledge.

Close reading data visualizations have the capacity to display one to two related texts by intralinear comparison. Readers of the New Testament have long performed such comparison through printed apparatuses putting related portions of the Synoptic Gospels together, and a similar data visualization is possible digitally. As the *digital Dante* project at Columbia University[9] shows, it is possible to compare Dante with Ovid in the original languages through the use of highlighting in yellow the lines of Ovid that inspired Dante and then including a caption that explains in English how the connection is relevant. Many more such projects could exist, particularly for texts such as biblical texts that have myriad texts from which they drew (e.g., Babylonian literature, Egyptian literature, Greek literature), myriad texts with which they shared a cultural milieu (e.g., early rabbinic texts, Greek and Roman texts), and myriad texts they inspired.

Variant readings demonstrate another area where visualization helps in the interpretation of the text. A visualization tool called *CollateX* in the VMR *(Virtual Manuscript Room)* brings together alternate attestations of the words of a text. In the *Codex Sinaiticus* Project, the manuscript and its transcription are displayed next to each other and suggested textual emendations may be found by hovering the cursor over highlighted text (Batovici & Verheyden 2021). As most scholars accept that there was a "Gospel of Mark" or a historical Jesus who uttered sayings, these attested words and suggested emendations are important because it is possible that they represent earlier and more authentic versions of the passage. For modern authors, variant readings are important, because there are authors such as Emily Dickinson who had variations

7 See http://www.perseus.tufts.edu (Accessed: 24 June 2024).
8 See https://copticscriptorium.org (Accessed: 24 June 2024).
9 See http://digitaldante.columbia.edu (Accessed: 24 June 2024).

within their own poems. In any event, variant readings preserve ancient and alternative interpretations of how a phrase could be imagined, and being able to see these even momentarily as part of the text instead of a footnote would be the ultimate goal of a reader. With practical limitations being what they are, variant readings in the collations provide a glimpse into a multiverse of interpretative potential.

1.2 Distant reading visualization

Distant readings of a text or multiple texts investigate detail patterns that might be missed by close reading. Writing of general literary studies and not in theological literary studies, Moretti posited that distant reading involved three separate domains: graphs for the linguistic details, maps for geographical associations, and trees for genres (Moretti 2005). I will apply this general tripartite structure in modified form to theological texts. In responding to comparisons of close and distant reading more than a decade after his original work, Moretti clarified that "attention to linguistic detail is common to both choices; all the rest is different" (Id. 2017, 688). Jänicke et al. frame the difference between close reading and distant reading as one of structure: "While close reading retains the ability to read the source text without dissolving its structure, distant reading does the exact opposite. It aims to generate an abstract view by shifting from observing textual content to visualizing global features of a single or of multiple text(s)" (Jänicke et al. 2017). Information visualizations have the advantage of providing statistical rigor to one's general impression that certain features of a text are recurring often, and they also have the ability to identify significant features of a large corpus that multiple readings and readers might have missed, whether a repetition of the word "one" or a greater preponderance of unnamed entities in some texts as opposed to other texts (Esposito 2022).

a. Graphs

Many projects within the realm of early Christianity have sought to provide what Digital Humanists have called distant readings of the corpora of early Christian texts in order to map common convergences.[10] One example of such a project is the *Paul and Patristics* project[11] of Jenn Strawbridge and Martin Hadley at the University of Oxford. Hadley organized Strawbridge's research into interactive tables, bar charts, and bubble matrix plots, all common forms of data visualization. The interactive table allows visitors to ask the database to display patristic authors by scriptural

10 For a definition of distant reading as pulling features of books out of their context to read against one another, see Hayles 2013.

11 See http://idn.it.ox.ac.uk/article/paul-and-patristics (Accessed: 24 June 2024).

references, region (Italy, Greece, Asia Minor, Egypt, North Africa, Palestine, Gaul, and unknown), and more. The bar chart is supposed to display references to authentic and disputed Pauline letters by century, early Christian author, and region; however, only the century option is consistently successful, and it highlights the predominance of Romans and 1 Corinthians diachronically. With the early Christian writer bubble matrix plot, one receives information by author about the author's Pauline references. Each graph stands on its own, and the information, though related, does not color one's view of the other graphs.

Another project that analyzes early Christian literature through charts is *Coptic Scriptorium.* The provided data visualizations focus on named/unnamed entities in a variety of corpora in world literature, ranging from Ruth to Cyrus to "English literature" to Shenoute to the New Testament. The data visualizations show that Coptic texts vary in named/unnamed entities but fall on the lower end of the spectrum of all world literature, with a traditional bar chart showing Shenoute below Mark and Besa below 1 Corinthians. This data visualization is particularly interesting, because it combines the computing interest in *named entity recognition* (cf. the chapter by E. Gius in this volume) with the theological interest in why some characters in texts appear unnamed and whether the lack of specificity indicates a lower status of the unnamed. With this data visualization, it is possible to see that the named and unnamed elements of the text might have meanings other than those attributed to them by textual scholars only reading for Jephthah's daughter, Rufus' mother, or Valens' wife. Such a possibility has been actualized in Schauf & Escobar Varela's (2018) network analysis that found female characters in the Javanese *wayang kulit* of the *Mahabharata* epic appear only sporadically and yet demonstrate high frequency of establishing bridges or betweenness in the story.

b. Maps

Maps are important supplements to textual analysis, because many of our religious texts mention multiple locales – often through travel – for the purposes of conveying the image of spreading the word extensively.[12] A major advantage of the digital turn of scholarship is the expanded capacities for the widespread dissemination of maps that it yields. As a visualization that typically relies heavily on color to convey meaning, a map is usually an expensive figure to include in a printed book, and its color possibilities are somewhat limited. With the progression of graphic user interface technology, it is now possible for maps to have a brilliant array of colors and to even allow for close-ups of the terrain on a platform like *Google Maps.* Not all time periods

12 As John Elsner says in analyzing the far-flung travels of Apollonius of Tyana, "[b]oth travels and miracles are instances of *thaumata,* the kinds of wonders which one would normally associate with a late antique holy man." (Elsner 1997, 23).

have equal access to the mapping possibilities, even for platforms such as Stanford's ORBIS that are built for Mediterranean antiquity.[13] Map interpretations still customarily limit one to a certain configuration of political geography that was extant for only a few years or decades. Coastlines have similar limitations, meaning that magnifying the terrain on *Google Earth* for every time period will simply not produce an accurate representation consistently as many ancient coastal cities have moved closer or farther from the sea due to silting and seismic activity (cf. the chapter by M. Wilkens in this volume).

Some of these problems may be viewed on non-scholarly interactive maps of Paul's journeys, which are the only maps the major search engine Google displays on the first few pages of results on a search for Pauline travels. *Google Maps* has a map of the journeys of Paul that has been created by simply pinning potential stops of Paul on a modern map. *Viz.Bible* is a project created by a sole Christian layperson, Robert Rouse. Unlike the *Google Maps* visualization, *Viz.Bible* has both an ancient map and a modern map. Four colors – green, purple, orange, and blue – indicate the potential routes of Paul. Yet, the main *interactive* component of this digitalization comes from the city links that lead to what should be information pages on a Classics site. Unfortunately, the information of the linked cities is not accessible, rendering the slider that changes the map from ancient to modern the only functioning interactive aspect of the visualization. Though *Viz.Bible* uses *OpenStreetMap,* which has some university affiliation through UCL *(University College London),* the two physical geography maps do not have a clear connection with one another. The lack of interactive mapping for figures from the New Testament, such as Paul, who have clearly delineated cities that they visit must be attributable to the scarcity of digital resources for biblical studies and the fear that technology is advancing too rapidly to make an interactive visualization viable for a sufficient amount of time to justify committing financial and temporal resources to such a project.

One workaround is to modify one's research question to fit the resources available by a broadly available humanities tool. For my undergraduate classes on early Judaism and Christianity, I had the students analyze the languages, locations, and purposes of Dead Sea Scrolls using Stanford's *Palladio* tool[14], which was developed for later historical periods. By interacting with the data themselves, students were able to understand the diversity of scrolls found at Qumran and in the Judean Desert, noticing which caves held which documents and whether there might be patterns. In addition, they also created word heat maps with readily available online tools and *R* from the so-called "sectarian scrolls," revealing a concern for concepts like the *Temple,* a concept that exist in physical form in Jerusalem.

13 See https://orbis.stanford.edu (Accessed: 24 June 2024).
14 See https://hdlab.stanford.edu/palladio (Accessed: 24 June 2024).

c. Trees

Moretti drew his distant-reading trees from evolutionary theory (Moretti 2003, 67), but trees as a data visualization tool for theology might also be seen in the linguistic treebanks associated with corpora, possibly influenced by Busa's use of the Porphyrian tree (Testori 2017). Many of these treebanks use *Universal Dependencies*[15], including the Latin treebanks: *Index Thomisticus Treebank* (ITTB); *Late Latin Charter Treebank* (LLCT); *Perseus; PROIEL; UDante.* Because of the diverse and diachronic usage of Latin, no single treebank can represent all theological Latin, stretching from earlier than the Vulgate to later than Vatican II (Cecchini et al. 2020). The treebanks of theological corpora, then, do not follow an evolutionary framework, but they point in the aggregate to the development of Latin as a language with a strong theological tendency. In exploring the treebanks, one might approximate the passage of theological time without having to slog through each minute itself as in close reading.

1.3 Summary

Data visualizations performed by modern scholars are able to bring together hundreds of manuscripts that were written over many centuries in a way never before possible. Access to manuscripts – provided they have been digitized and authorized – is radically democratized. In such a way, it is easier to compare manuscripts to one another and to see the possibilities for diachronic and synchronic comparison. Close reading is helped by visualizations of other occurrences of a word, morphology, syntax, definition(s), variants, emendations, cultural background, and even geography. Distant readings help create graphs of character and textual relations, maps arising from manuscript sources, and trees of indices. All in all, visualization of theological texts serves to enhance the exegetical reading of texts and corresponding understandings of how texts fit into the broader literary, theological, social, and political landscape.

2. Problems with information visualization

Although visualizations of theological texts are inherently useful, their creation is fraught with potential problems. These range from creating the impression that numerical frequency is what constitutes religious feeling in a text to detracting from the beauty of the text with a distractingly aesthetically pleasing chart. Some of our data simply has not found its way into our data sets due to a lack of sufficient resources

15 See https://universaldependencies.org (Accessed: 24 June 2024).

to digitize all available liturgical and other theological texts, and visualizations of the current categories might reinforce the neglect of theological texts outside of liturgical scrolls and codices, such as *mezuzot,* door lintel inscriptions, and other texts placed on practical objects. While these problems do not seem of a magnitude to preclude visualizations of theological texts, it is important to be aware of opportunities to better envision the textual world.

Aesthetics are an important part of visualization charts, maps, and trees, and they have the potential to distract from the message of the information. Though most data visualizations are produced digitally, they are expected to resonate artistically and occasionally have hand-drawn aesthetics. This means that data visualizations can inspire feeling in a way that is potentially unconnected with the data the visualizations interpret (Brinch 2020, 259f.; Simpson 2020,164). As many scholars have noted, it is possible for exceptionally "beautiful visual renderings" to signal authority and accuracy where none exists (Laaksonen & Pääkkönen 2020, 104.). In instances where the data visualization is of a negative phenomenon, such as the spread of tuberculosis evidenced by parish records, there is the potential for the beautiful aesthetics of the data visualization to dull the depth of the anguish felt by individuals and their social circles without some complementary information such as personal narratives (Nygren et al. 2016). On the other hand, it would seem that some readers have become so accustomed to data visualizations that readers do not interact with them. A study of a *New York Times* website identified that only 10–15 percent of readers attempted to absorb information in its interactive data visualizations, potentially signaling the data visualization is accepted as an unquestioned authority for a story (Ridley & Birchall 2020, 132). Literary critics have also pointed out that the rationale for close reading is to find the beauty in the text, which seems to be an enterprise that is not at all present in information visualizations, which satisfy themselves with being aesthetically pleasing in and of themselves instead of pointing the way to innate beautiful texts, whether theologically or secularly literary ones (Aquilina 2017, 513).

In processing the texts, different subspecialties of theology might encounter different problems and certain theological authors might have a way of framing linguistic conventions that diverge from other contemporary linguistic practice. Chinese and Arabic, for instance, require more preprocessing for tokenization than languages such as English, requiring their own segmenter.[16] While the concept of *presence* in Latin can take on very mundane meanings, Busa began his Index specifically because he wanted to account for the connection in Aquinas' thought of *praesens* with the preposition *in.*[17] This would complicate visualizations, because not all word heat maps or word clouds with *praesens* would accurately convey the nuance found in Aquinas or another instance where tokenization becomes crucially important.

16 See https://nlp.stanford.edu/software/segmenter.html (Accessed: 24 June 2024).

17 According to Busa, "I soon learned that such words in Thomas Aquinas are peripheral: his doctrine of presence is linked with the preposition *in*" (Busa 1980, 83).

Another potential concern of information visualization is the perception that statistical frequency automatically yields greater textual and social significance. A class data visualization project on Islamic sources run by Jeff Blecher at George Washington University demonstrates several areas where number does not automatically equal the correct and singular response. If sources attested multiple spellings for the same city (e. g., Medina, Medinah, and Madinah), would a standard spelling be necessary and would the standard spelling have to be the most attested? If listed occupations range from judge to scholar, would the most frequent automatically be the most powerful? How would individuals with multiple occupations divide their importance (Blecher 2016, 238)? Scholarly studies into the Septuagint find a similar diversity in that the LXX treats Hebrew *hapax legomena* in a variety of ways: omission, transliteration, association with similar-looking words, and so forth (Verbeke 2008; 2011). Information visualizations relying on statistical frequency lose the richness of *hapax* words and their transference into various texts. Pilate's wife, for instance, might only feature in one of the four canonical gospels, but that does not make her action of unsuccessfully attempting to warn her husband just as Calpurnia was thought to have tried to warn Caesar any less portentous.

Likewise, information visualizations have to contend with very sporadic manuscript availability, with certain regions and languages heavily privileged over others due to historical and climactic circumstances. In order to accommodate manuscript collections and academic training patterns, projects form around texts, textual traditions, languages, regions, and manuscript dates. Choices have to be made: how should we organize the scant information we have? One answer is from the item generating system at the *Trismegistos* Project[18], which creates items based on the material manuscript. With *Codex Alexandrinus,* there is an example of how a rarely found text in manuscripts and reception history *(1 Clement)* can have an early instance of it being included with the canonical New Testament. In projects such as that of *Trismegistos,* the material fact of the text's inclusion may be accounted for without recourse to other instances of a New Testament sans *1 Clement.* Thus, TM 62318 is linked to British Library images of *Codex Alexandrinus* and described as having the Old Testament, the New Testament, Athanasius, and Clement of Rome despite missing some verses of *1 Clement.* Allowing manuscripts to retain the particular circumstances of their creation and re-finding gives a more historically accurate and less anachronistic version of history. Another answer is that texts might be virtually reunited, whether in the sense of rejoining a manuscript that has its fragments scattered across the globe or in the sense of bringing together a number of witnesses to a sketchily present text (Miyagawa 2021, 181; Wolf et al. 2011).

Durability represents a further consideration of information visualization. Planned obsolescence is necessary for Digital Humanities projects. Projects require ongoing physical resources (e. g., servers, storage, electricity), and they may encounter

18 S. https://www.trismegistos.org (Accessed: 24 June 2024).

problems with links to external sites that change or with applications that are no longer supported. While it may be beneficial to students in a particular course to use a contemporary business trend such as *R, Python,* or *SQL* to create data visualizations, durability of a project will be assured by translating true humanities projects – such as the *Index Thomisticus* – to the digital platform and letting those inform data science rather than letting data science set the parameters of what constitutes a *good* project in data visualization. Just as IBM and punch cards are no longer as cutting edge as they once were, information visualization that is too *of the moment* will be restricted to that moment as technological advancement continues along the lines of Moore's Law or whatever law replaces it. Transferring digital information visualization back to the analog sphere via paper publishing might not only help to preserve the project longer, but it would also save electricity and carbon credits. Recyclable transcriptions and providing a "what you see is what you get" environment in which to produce major editions rather than forcing everyone into XML or its analogue should be viewed as a best practice (Houghton 2013).

3. Comparing data visualization in other fields to theology

Caveats aside, data visualization held a place in academic circles prior to the *Index Thomisticus,* and it represents a useful way to engage readers in the interpretation of theological texts. Information visualization will only continue to become more sophisticated in terms of metadata analysis and visual display. Some practices in academic data visualization would benefit data visualization in theological and religious studies. Among these are research maps and augmented and virtual reality projects.

In science, the idea of a *Map of Science* as initiated by UCSD and continued by Georgetown helps to analyze researchers, projects, and funders of various areas of science.[19] For UCSD, the *Map of Science* needed to bring together two sets of publishing data (Börner et al. 2012). The analyses performed with these maps are perhaps imprecise. Various scientific papers have to be incorporated in the dataset by title, keywords, and abstracts regardless of the academic departments of the researchers, leading to the possibility that some articles might be missing and some articles might be duplicated. Research clusters of 50 to 15,000 form when researchers cite one another in papers, leading to the possibility that new fields and subfields might be created. When gender, institutional affiliation, country, ethnicity, and other categories are known, analyses might be run about the strengths of various categories in various traditional and emerging fields of sciences. Researchers might find collaborators working on the same topics by consulting the latest iteration of the *Map of Science.*

19 See https://sciencemap.eto.tech (Accessed: 24 June 2024).

A comparable capability to visualize the scholarly intersections of the field could benefit theological studies. It would be possible for researchers on different biblical themes in different disciplines to find each other and collaborate. Areas of theology neglected regionally could be addressed. Potential gender gaps in areas of research could be found. Trends in established subdisciplines could be identified, and emerging subdisciplines could be discovered. As a Catholic biblical scholar, I have mostly been trained in biblical studies by Jewish, Protestant, Catholic, and atheist biblical scholars, but I have also used my Catholic education to publish on biblical reception and theological topics in the history of Catholicism. One wonders at how subdisciplines, religious traditions, theological training, and theological affiliation would intersect and diverge in a visualization of theological scholars and their research.

In addition to a map of theological studies that lays out subdisciplines and affiliations, another idea in the broader realm of digital data visualizations that would be helpful for theological studies is that of augmented reality and virtual reality. In the current digital landscape, there are scholarly virtual recreations of salient moments in religious history. One such project is the *Virtual Paul's Cross* Project, which takes its audience into a reconstruction of the church setting as John Donne gives a sermon to support James in his overcoming of the Gunpowder Plot. By developing the sounds of the audience and other features of the text, Wall has come to see the sermon less as a "text" than as a "trace" (Wall 2014). Our theological texts, however scriptural they may be, often existed originally primarily as speech in larger theological events. By being able to click from a description of a meeting between Peter and Paul in Jerusalem to a possible setting of such an encounter with its accoutrements based on archaeology of first-century Jerusalem, a layer of interpretive potential would be added to close reading as powerful as having other instances of a particular lexical choice attested by a majority of textual witnesses. Alternatively, were we situating our reading in a particular tome such as *Codex Alexandrinus*, we could have a visualization of a copying room in the fifth century C.E. Virtual and augmented reality projects are essentially multidimensional maps at close range, and they aid one in immersing oneself in circumstances that are no longer present.

4. Conclusion

Information visualization is an important tool for interpreting theological textual sources. Visualizing words in theological texts can improve their close reading through linguistic aids and literary-historical contextualization and enable their distant reading, whether in graph, map, or tree formats. In terms of the Digital Humanities, the entire field started auspiciously with the indexing of the complete works of Thomas Aquinas by Father Roberto Busa's scholarly team and the donated computing resources of IBM. Thus, in the sense that an index is a table visualizing word frequency in corpora, digital data visualization too might be traced back to the late 1940s and 1950s. To claim that information visualization is particularly innovative, unknown, or threatening to existing ways of examining texts humanistically is hyperbolic. Data visualization enriches our understanding of concordances, a tool of theological study since at least the Middle Ages, and how they help us read our texts. Visualization allows us to see when authors are invoking a particular theme, text, or author repeatedly and diachronically. Language patterns across a collection of texts in a library can be better discerned through it.

There are several potential problems with information visualization in theological studies, which are typically found elsewhere. Visualizations might become in an end unto themselves, losing sight of the delight they were trying to convey in the first place. Durability becomes an issue when technology advances or external sites reorganize the content to which visualizations were linked. In certain cases, an emphasis on statistical frequency diminishes the emotional impact of intense events upon individual lives and their texts. Tokenization, lemmatization, and other preprocessing issues might distort visualizations in terms of word frequency or definition. The categories of the visualization might preclude other valid interpretations, such as the choice between reading a manuscript as a localized document in its own right or as an ancient and somewhat universal collation of earlier manuscripts.

All in all, though, information visualization has much to offer with these considerations taken into account. Theological scholars broadly should have the opportunity with democratized access to conduct visualization experiments to see what patterns emerge. Funding sources should follow the example of IBM in letting scholars of theology and the humanities to lead the way and determine the course and goals of the project, so that theology and the humanities may humanize the development of data visualization happening in other sectors of knowledge creation. To this end, should theological scholars look farther afield to see what aspect of data visualization they might bring into their discipline, the *Map of Science* – which brings together researchers in different subfields by charting connections between researchers – would be a helpful model. Theological data visualization projects might continue to inspire each other, whether in making treebanks of an already extant index or in creating new virtual and augmented reality projects that allow scholars to visualize the settings in which the text was first created and rehearsed along with the word frequency, authorial allusions, and so forth of the texts within given corpora. Although the process of information visualization might seem a little contrived and susceptible to oversimplifying data, one must remember and take inspiration from the verbal data visualization of the creation of world at the beginning of the Book of Genesis – and, indeed, the Bible – in its organization of creation into seven simple categories of light, sky, land, plants, animals, people, and rest and in the canonical need to provide yet another interpretation of the same event with a different data set.

References

Aquilina, M. (2017). The Work of the Literary Critic in the Age of Big Data, *Interdisciplinary Literary Studies,* 19(4), 493–516.

Batovici, D., & Verheyden, J. (2021). Digitizing the Ancient Versions of the Apostolic Fathers. Preliminary Considerations. In T. Hutchings & C. Clivaz (Eds.), *Digital Humanities and Christianity. An Introduction* (pp. 103–124). Berlin/Boston: De Gruyter.

Blecher, J. (2016). Pedagogy and the Digital Humanities. Undergraduate Exploration into the Transmitters of Early Islamic Law. In E. Muhanna (Ed.), *The Digital Humanities and Islamic and Middle East Studies* (pp. 233–250). Berlin/Boston: De Gruyter.

Booth, A. (2017). Mid-Range Reading. Not a Manifesto, *Publications of the Modern Language Association of America,* 132(3), 620–627. URL: https://www.jstor.org/stable/27037375 (Accessed: 24 June 2024).

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., Larivière, V., & Boyack, K. W. (2012). Design and Update of a Classification System. The UCSD Map of Science, *PLoS ONE,* 7. DOI: https://doi.org/10.1371/journal.pone.0039464 (Accessed: 24 June 2024).

Brinch, S. (2020). What we talk about when we talk about beautiful data visualizations. In M. Engebretsen & H. Kennedy (Eds.), *Data Visualization in Society* (pp. 259–276). Amsterdam: Amsterdam University Press, 2020).

Burton, D. M. (1984). Review of *Index Thomisticus. Sancti Thomae Aquinatis operum indices et concordantiae* by Roberto Busa. *Sancti Thomae Aquinatis opera omnia* by Thomas Aquinas and Roberto Busa, *Speculum,* 59(4), 891–894.

Busa, R. (1949). Announcements, *Speculum* 25, 424 f.

Id. (1980) The Annals of Humanities Computing. The Index Thomisticus, *Computers and the Humanities,* 14, 83–90. URL: https://www.jstor.org/stable/30207304 (Accessed: 24 June 2024).

Cecchini, F. M., Korkiakangas, T., & Passarotti, M. (2020). A New Latin Treebank for Universal Dependencies. Charters between Ancient Latin and Romance Languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 933–942). Marseille: ELRA. URL: https://aclanthology.org/2020.lrec-1.117 (Accessed: 24 June 2024).

Collins, J. (1951). Review of *La terminologia tomistica dell'interiorità* von R. Busa, *Speculum,* 26, 495–497.

Elsner, J. (1997). Hagiographic geography. Travel and Allegory in the Life of Apollonius of Tyana, *Journal of Hellenic Studies,* 117, 22–37. DOI: https://doi.org/10.2307/632548 (Accessed: 24 June 2024).

Esposito, E. (2022). Reading Images. Visualization and Interpretation in Digital Text Analysis, *Artificial Communication. How Algorithms produce Social Intelligence*

[Blog]. Cambridge, Mass.: MIT Press. URL: https://artificialcommunication.mitpress.mit.edu/pub/jtkikw5z (Accessed: 24 June 2024).

Hayles, N.K. (2013). Combining Close Reading and Distant Reading. Jonathan Safran Foer's 'Tree of Codes' and the Aesthetic of Bookishness, *Publications of the Modern Language Association of America,* 128(1), 226–231.

Houghton, H.A.G. (2014). The Electronic Scriptorium. Markup for New Testament Manuscripts. In C. Clivaz, A. Gregory & D. Hamidovic (Eds.), *Digital Humanities in Biblical, Early Jewish and Early Christian Studies* (pp. 31–60). Leiden: Brill [= *Scholarly Communication,* 2].

Jänicke, S., Franzini, G., Cheema, M.F., & Scheuermann, G. (2017). Visual Text Analysis in Digital Humanities, *Computer Graphics Forum,* 36, 226–250.

Jones, M., & Yeandle, D. (2021). Obiturary. Professor Roy Albert Wisbey (1929–2020), *German Life and Letters,* 74, 161–165.

Laaksonen, S.-M., & Pääkkönen, J. (2020). Between Automation and Interpretation. Using Data Visualization in Social Media Analytics Companies. In M. Engebretsen & H. Kennedy (Eds.), *Data Visualization in Society* (pp. 95–110). Amsterdam: Amsterdam University Press.

Miyagawa, S. (2022). *Shenoute, Besa and the Bible. Digital Text Reuse Analysis of Selected Monastic Writings from Egypt* [Diss.]. Philosophische Fakultät: Georg-August-Universität Göttingen. URL: https://ediss.uni-goettingen.de/handle/11858/13894 (Accessed: 24 June 2024).

Moretti, F. (2003). Graphs, Maps, Trees. 1, *New Left Review,* 24, no pag. URL: https://newleftreview.org/issues/ii24/articles/franco-moretti-graphs-maps-trees-1 (Accessed: 24 June 2024).

Id. (2005). *Graphs, Maps, Trees. Abstract Models for Literary History.* London: Verso.

Id. (2017). A Response, *Publications of the Modern Language Association of America,* 132, 686–689.

Nygren, Th., Frank, Z., Bauch, N., & Steiner, E. (2016). Connecting the Past. Opportunities and Challenges in Digital History. In M. Hayler & G. Griffin (Eds.), *Research Methods for Creating and Curating Data in the Digital Humanities* (pp. 62–86). Edinburgh: Edinburgh University Press.

Passarotti, M. (2019). The Project of the Index Thomisticus Treebank. In M. Berti (Ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution* (pp. 299–320). Berlin/Boston: De Gruyter Saur [= *Age of Access? Grundfragen der Informationsgesellschaft,* 10]. DOI: https://doi.org/10.1515/9783110599572-017 (Accessed: 24 June 2024).

Ridley, A.L., & Birchall, Ch. (2020). Evaluating Data Visualization: Broadening the Measurements of Success. In M. Engebretsen & H. Kennedy (Eds.), *Data Visualization in Society* (pp. 127–140). Amsterdam: Amsterdam University Press.

Schauf, A., & Escobar Varela, M. (2018). Searching for Hidden Bridges in Co-Occurrence Networks from Javanese *Wayang Kulit, Journal of Historical Network Research,* 2, 26–52.

Simpson, J. (2020). Visualizing data. A lived experience. In M. Engebretsen & H. Kennedy (Eds.), *Data Visualization in Society* (pp. 157–168). Amsterdam: Amsterdam University Press.
Szondi, P. (1978). Introduction to Literary Hermeneutics, *New Literary History,* 10(1), 17–29. DOI: https://doi.org/10.2307/468303 (Accessed: 24 June 2024).
Testori, M. (2017). Methods of Quality, Quality of Methods. What does Roberto Busa have to Communicate to Digital Humanists in the 21st Century? From Hermeneutics to Performativity, *Digital Humanities Quarterly,* 11(3), 1–9. URL: http://www.digitalhumanities.org/dhq/vol/11/3/000329/000329.html (Accessed: 24 June 2024).
Tufte, E. R. (1983). *The Visual Display of Quantitative Information.* Chesire, CT: Graphics Press.
Uriel, S., & Greenstein, E. L. (1988). The Religious Significance of the Peshat, *Tradition. A Journal of Orthodox Jewish Thought,* 23, 41–63.
Verbeke, E. (2008). The use of Hebrew hapax legomena in Septuagint studies. Preliminary remarks on methodology. In H. Ausloos, B. Lemmelijn & M. Vervenne (Eds.), *Florilegium Lovaniense. Studies in Septuagint and trextual criticism in honour of Florentino García Martínez* (pp. 507–521). Leuven: Peeters.
Ead. (2011). *Hebrew hapax legomena and their Greek rendering in LXX Job* [Diss.]. Faculteit Theologie en Religiewetenschappen: KU Leuven. URL: https://lirias.kuleuven.be/1875994 &lang=en (Accessed: 24 June 2024).
Wall, J. N. (2014). Transforming the Object of our Study. The Early Modern Sermon and the Virtual Paul's Cross Project, *Journal of Digital Humanities,* 3(1). URL: https://journalofdigitalhumanities.org/3-1/transforming-the-object-of-our-study-by-john-n-wall (Accessed: 24 June 2024).
Wolf, L., Littman, R., Mayer, N., German, T., Derschowitz, N., Shweka, R., & Choueka, Y. (2011). Identifying Join Candidates in the Cairo Genizah, *International Journal of Computer Vision,* 94, 118–135. DOI: https://doi.org/10.1007/s11263-010-0389-8 (Accessed: 24 June 2024).

Digital Edition

Annette von Stockhausen

https://orcid.org/0000-0001-5382-6322

Abstract The article first explains the characteristics of digital editions and then discusses the requirements that must be met by a digital edition ("FAIR principles", structural requirements). In a further section, the potential of a digital edition compared to the traditional print edition is presented (interoperability, different usage scenarios, further development, transparency). Finally, the current limitations for both the productive and the receptive use of digital editions are addressed: lack of expertise, lack of tools, necessary interdisciplinary cooperation, pending standardization and recognition of digital editions as scientific achievements.*

Keywords Digital Edition, Presentation, FAIR Principles, TEI, Publishing

1. What is a Digital Edition?

Like its analog, printed counterpart, the digital scholarly edition provides basic access to sources for various areas of theology,[1] particularly areas doing historical research.[2] Digital editions make historical texts accessible and present them using rules initially defined or, in most cases, standardized long ago and then applied.[3] Depending on the subject of the edition and the editorial practice prevalent in the respective partial/ specialist discipline (critical, diplomatic/documentary, genetic), an idealistic or materialistic concept of the text is taken as a basis.[4] The digital edition is then produced so

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

1 For a foundational volume on Digital Editions, see Apollon et al. (2014); Pierazzo (2015); Driscoll & Pierazzo, (2016); Boot, et al. (2016) and Bleier et al. (2018).

2 The focus is primarily on subjects that deal with the Old Testament, the New Testament, and the history of Christianity. The spectrum of sources and the languages, materials, and contexts that transmit these sources is correspondingly broad and requires adapted methods in each case.

3 The following remarks focus on textual sources; however, most of what is said below also applies to oral or musical sources.

4 Between the two extremes of an idealistic (edition of works) or materialistic concept of text (edition of individual manuscripts, inscriptions, or historical prints) there are, of course, many nuances that can make digital editions visible in comparison to printed editions. From an ancient studies perspective, see, e.g., Meins (2016).

Annette von Stockhausen: Digital Edition, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 333–344. DOI: https://doi.org/10.11588/heibooks.1521.c21957

that the edition chooses an editorial method and form of presentation appropriate to its subject and tradition (cf. Sahle 2014).

The printed edition presents the edited text as the result of editorial, selective, and/or normalizing work, usually mediated by a publisher and potentially as part of a chronologically and/or thematically oriented series. In addition to the presentation of the edited text, an edition also documents (comprehensively) both the transmission findings or the transmission carriers and the editorial decisions, since only in doing so can the quality and reliability of the edition be assessed by the recipient; otherwise, users would have to refer to the sources on which the edition is based. Moreover, the digital edition is made accessible by means of various indices to enable selective access to the edited text based on questions of content in addition to the generally linear reading. However, materials such as transcriptions or collations (in tables or other forms), which were created in the course of the editorial work, are not made accessible to the recipients, or at best in a highly selective manner.

If the printed edition can therefore be described primarily in terms of its form of presentation, the digital edition, on the other hand, is fundamentally characterized by the fact that there is a strict separation between the data stored digitally in a specific file format (transcriptions, collations, edition, etc.)[5] and its digital (as a website, e-book) or analog (as a book) form. For a human user, however, this strict separation is often not recognizable at first glance, because one usually encounters the digital edition first or even exclusively at the level of presentation.

There is a characteristic separation of content (data) and form (presentation) unfamiliar within the book paradigm, where the two are inseparably fused.[6] The data of a digital edition can manifest itself in different forms of presentation and display, which corresponds to very different needs and research questions and covers only partial aspects of the data. By detaching the data from its *(one)* presentation,[7] a digital edition is thus open to very different perspectives, whereby the editor will normally present one's perspective(s) together with the publication of the data. The edition *can* then be received in the way presented but *must* not be, given the recipient has the

5 The de facto standard is the *Extensible Markup Language* (XML): https://www.w3.org/XML (Accessed: 18 June 2024) in the standardizing of the *Text Encoding Initiative* (TEI): https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html (Accessed: 18 June 2024). However, theoretically (and in certain cases, practically) there are other conceivable formats, such as, SQL or graph databases. Ideally, the data as the "single source of truth" is the basis of all forms of presentation of a digital edition.

6 A comparable case would be the creation of an *editio minor* alongside an *editio maior* or that of a popular/reading edition alongside the scholarly edition. In the case of the digital edition, however, this fusing is created programmatically from the data of one edition (see previous note), whereas in a book, at least in the pre-digital age, two different print templates were created.

7 In principle, a TEI XML file is also a form of presentation, albeit a very particular one for non-technical users.

option of evaluating the data of the edition in a different way and under different premises, annotating and/or presenting the data further.[8]

The separation of data and its presentation has a further consequence. A printed edition is designed to be received by a human reader, whereas a digital edition, in its data form, is suitable for reception (and further processing) by a computer or computer programs (i.e., machine-readable), whether via an *application programming interface* (API) or a download. However, in the best-case scenario, the edition can (and should) also remain human-readable beyond the presentation (i.e., *the graphical user interface*, GUI) in the sense that a TEI-XML file itself is already a form of presentation.

The data is therefore the most important aspect of a digital edition. The data model behind the edition and the format in which it is stored therefore are critical. Standardization is already advanced. XML, which follows the guidelines of the *Text Encoding Initiative* (TEI), has *de facto* established itself as the data format, despite individual peculiarities of the XML format and the fact that TEI is basically not *one* language, but rather appears in a multitude of "dialects" not readily understandable to each other, have received criticism.[9]

What exactly is a digital edition: the data *or* the presentation of the data or the data *and* its presentation?[10] In the theoretical discussion, this question is usually treated on the level of "data" vs. "data and presentation," yet in practice the question is often answered in terms of presentation, insofar as the data is not made available to the recipient at all. How one answers this question also depends on whether, with Patrick Sahle, one understands the essential characteristic of a digital edition in the fact that a digital edition is digital in theory, method, and practice,[11] and therefore cannot be transferred to print without a loss of information or functionality,[12] because doing so would manifest itself in the presentation layer, which is expressed in the structures and peculiarities of the internet (keyword: linking). However, one must then necessarily discuss how the loss of information and functionality should be quantified

8 Doing so requires editors to "let go" and "tolerate" other interpretations and presentations in a way unfamiliar from the print paradigm, and users to be able to process the data themselves. See also below (Section 4.1.).

9 Criticism is mainly directed at the tree structure of XML and the resulting difficulties in modeling overlaps (https://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html). Cf. Cummings (2018). Proposed alternatives are database-based, such as using graph databases, esp. A. Kuczera, e.g. Neill & Kuczera (2019), or SQL databases, see Cadmus: https://myrmex.github.io/overview/cadmus (Accessed: 18 June 2024) by Daniele Fusi. A more recent approach is the combination of TEI semantics and graph databases, as proposed by Kuczera (2022).

10 On this distinction, see Barabucci et al. (2017). The answer to the question is usually connected with a further aspect of digital editions, namely, long-term archiving; more on this idea later.

11 Cf. Sahle (2016, 28): "Scholarly digital editions are scholarly editions that are guided by a digital paradigm in their theory, method and practice."

12 Cf. ibid., 27: "A digital edition cannot be given in print without significant loss of content and functionality." In the article, Sahle draws on the extensive explanations in his dissertation (Sahle 2013a; b; c). This definition is highly influenctial in the current discussions.

and where the boundaries are. Another related aspect is the long-term availability or long-term archiving: particularly when an edition is "digital," according to Sahle's definition, which arises primarily at the level of presentation and less so at the level of data. If the presentation on the internet is up to date in terms of technical possibilities, then in view of the speed of development, a presentation must be updated to the current programming status after a few years at the latest, i. e., the programming must be developed further (but can also be equipped with new functions that were not previously feasible in view of new possibilities). The internet is constantly changing and evolving, and presentations must inevitably follow suit and be maintained permanently with greater effort than a book. The situation is different with the data of an edition, as the standards used develop slower and ultimately more generically,[13] so that the data can be updated with comparatively less effort for new versions. For this reason, too, the data should be seen as the core of a digital edition, on which all possible presentations are based.

In his development of a definition, Sahle introduces a further significant differentiation,[14] esp. in current practice, and must therefore be mentioned: the distinction between digital and digitized editions.[15] The digitized edition is mainly found where old print editions are retro-digitized, but sometimes also in new editions, when one simply transfers the print edition into the digital medium, i. e., to give it a digital form,[16] without exploiting the additional possibilities associated with digitality (see below in chap. 3). According to Sahle, the digitized edition is defective (and not to be desired); however, the digitized edition nevertheless remains important in the context of retro-digitization, because it allows the printed body of knowledge (and editions) to be transferred into the digital age and thus made usable and researchable using computerized methods. One need not expect that new, digital editions of all texts that are important for research in the respective subject area will be produced in the foreseeable future, or at all. In this sense, the digitization of printed editions represents a break comparable to the transfer of texts from manuscript to print and

13 TEI P5 was published in the first version in 2007, see https://doi.org/10.5281/zenodo.3556213 (Accessed: 18 June 2024).

14 See Sahle 2016, 27: "A digitized edition is not a digital edition." Cf. ibid., 33.

15 "Digitized" does not mean a book is scanned and then made available as a PDF or graphic file; rather, it means that the information is made available as a full text that can be processed further by a computer. In principle, the diverse full-text databases, such as the *Thesaurus Linguae Graecae,* https://stephanus.tlg.uci.edu (Accessed: 18 June 2024), offer digitized editions, yet without the scholarly apparatus fundamental to an edition.

16 The guidelines of the *Library of Digital Latin Texts,* see https://github.com/DigitalLatin/guidelines (Accessed: 18 June 2024), has some traits of digitization, cf. already Stockhausen (2020) 124, note 10. However, digitization can also be found in unexpexted places, such as the TEI guidelines, where the relevant section of the guidelines, see https://tei-c.org/release/doc/tei-p5-doc/en/html/TC.html (Accessed: 18 June 2024), is called "Critical Apparatus" rather than something like "Textual Variation," using terminology of the critical print edition.

potentially has similar consequences for contemporary research as this type of transition had for scholars of the early modern period.

2. What is needed for a Digital Edition?

2.1 "FAIR" – Findable, Accessible, Interoperable, Reusable

To fulfill their potential, digital editions must be implemented according to the "FAIR principles,"[17] i. e., they must be findable, accessible, interoperable, and reusable. Ideally, digital editions are available as individual texts and, where applicable, as a corpus in open access or as open data under an open license via API (programming interface) and/or as a download, given that most of the FAIR principles can be implemented under these conditions. For digital editions (as data and/or in their presentation form) to be used and evaluated, they must be addressable and able to be bibliographed, preferably in such a way that they are permanently accessible using *one* link and able to be cited with the same web address,[18] i. e., have permalinks, the permanence of which must be ensured by the providing institution (academic organization, library, publisher). Permanent accessibility is the greatest challenge of digital editions beyond the actual editing work.[19] As a result, digital editions not only consist of the edition data itself, but also include *meta*data for the data and for the various subsequent uses and forms of presentation. This metadata ensures clarity for the recipients (whether human or computer) regarding the contents of the edition, to what extent, in which format, under which conditions, and by whom the edition was created and where it can be found.[20] In addition to the metadata, the editorial standards and (individual) decisions as well as the technical workflows and solutions should be documented as comprehensively as possible.[21]

17 See https://www.go-fair.org/fair-principles (Accessed: 18 June 2024). Cf. Wilkinson et al. (2016).

18 In competition with printed editions, the issue of long-term availability and citation is fraught with considerable difficulties, as digital editions should in principle be available permanently, spanning decades and centuries, just like printed editions, but the actual experience with them has been the opposite, because digital projects are often no longer available very soon after the end of the project and the expiration of funding. In any case, long-term availability means long-term maintenance to an extent that far exceeds what is necessary for printed material.

19 Of course, printed editions (and their sources) have also been lost due to natural disasters or human influence, and their existence is known only due to their metadata.

20 Standardized vocabulary, such as Dublin Core, see https://www.dublincore.org/specifications/dublin-core (Accessed: 18 June 2024) or other library standards, should also be used for metadata.

21 There remains a lack of standards in this regard.

Since digital editions can be corrected and updated much more easily due to their digital nature (see below in chap. 3), i.e., new editions can occur much more frequently than in print, versioning is another important aspect to consider.[22]

2.2 Structural Requirements

Standards should be followed in all aspects of the edition. Doing so guarantees interoperability and enables subsequent use (and preferably established and not outdated). Individual solutions,[23] especially if they are not documented, and innovation for innovation's sake,[24] should therefore be avoided wherever possible.

Standardization is even more important given digital editions involve much more effort than printed editions,[25] even in the absence of well-established processes and involved parties, and a large part of this additional effort cannot simply be transferred to other parties such as publishers (layout, production, distribution) or libraries (metadata, citation, long-term availability), at least not yet. Publishers – including those relevant for theological publications – currently offer neither the necessary editing and publication workflows nor the technical infrastructure;[26] rather, they offer digitized books in PDF format as the standard (cf. Arnold & Döhnert 2024). At the same

22 Versioning is best based on a Version Control System (VCS), as doing so also makes tracking changes easier.

23 A compromise must be found between the best possible modeling of one's own edition and the greatest possible connectivity to editions that are thematically or chronologically similar or based on comparable sources. Although it may be or seem to be more often in the interest of third-party funders, it should also be considered whether one's own edition project can be published not only on a separate (project) website, but also in a corpus of similar digital editions, since doing so also facilitates subsequent use.

24 Innovation for innovation's sake is a result of digitality (in addition to the practice of current research funding), claiming innovation is easier to achieve and is in demand; however, innovation is a hindrance to the aspect of long-term availability and reusability, because completely new technologies may be used that cannot establish themselves and are therefore not supported or that may require so much technical infrastructure that they cannot be provided permanently by the sponsors.

25 However, the additional work is not limited to publication and ensuring permanent usability. The work also extends to the actual editorial work if digital editions are to offer *more* than their printed counterparts, which ultimately also contain *more* information and therefore require more effort.

26 One exception thus far is Brill, which has set up its own publication platform for digital editions based on the *Scaife Viewer* (https://scaife.perseus.org) with *Brill's Scholarly Editions* (https://scholarlyeditions.brill.com). Both addresses were accessed on 18 June 2024. However, even in this case, only retro-digitized editions and no *born digital* new editions are available; further development remains to be seen in view of the fee-based offer, especially regarding the aspects of interoperability and reusability. It is becoming apparent that all major academic publishers that publish printed editions will develop concepts and solutions for digital editions in the foreseeable future.

time, digital editions as a special type of publication are part of the wider context of the current discussion about *open science* and *open access.* These discussions are also part of a larger shift in publication structures away from commercial publishers and towards libraries and universities or academic organizations' own publishing initiatives, including infrastructures that go beyond pure publication.[27] In addition, there are specialized service providers who have the necessary technical capabilities, e.g., for XML and its further processing, databases, web/application programming or user experience (UX). Nevertheless, the responsibility of the individual editing project currently remains great and is best solved collaboratively. The requirements for technical skills, software, infrastructure, and distribution are complex and require ongoing maintenance, which not only includes the necessary updates, but also the provision of server infrastructure and its financing.

3. Opportunities Presented by Digital Editions

If digital editions implement the "FAIR Principles" and are published in *open access* and as *open data,* then there are new possibilities created compared to print editions.[28] They are – provided the user has internet access[29] – available worldwide and can be used by anyone. Digital tools, either those offered by the editors in the presentation layer of the edition itself or applied to the data by the user independently, enable access not only for specialists or a wider specialist audience, but for anyone interested.

Digital editions can accommodate different user scenarios and capabilities by allowing users to interactively show or hide different functionalities or edition layers according to their own preferences. The text of the edition can be searched and analyzed in a variety of ways – ultimately limited only by the functionalities of existing tools or the user's programming skills. Comments, annotations of entities, and other enrichments of various kinds, such as the integration of digital facsimiles and links to other offers on the internet such as databases, dictionaries, encyclopedias, or editions (including digitized books) offer potentially unlimited information. Users can also further enrich this information with their own annotations and additional links, either in the original edition context, provided the editors enable or allow additions

27 Libraries like the Herzog August Library in Wolfenbüttel, the Darmstadt University and State Library, and the Heidelberg University Library are already playing an important infrastructural and editorial role in digital editions of medieval, early modern, and modern texts.

28 If digital editions do not comply with the FAIR principles and/or are not published in *open access* or as *open data,* then they may still offer *more* than printed editions, e.g., a full-text search ability, but, at the same time, there is a risk that long-term availability is not guaranteed.

29 In this respect, digital editions are again limited, whereby not only the availability of internet access, but also censorship must be considered. On the other hand, the usability of printed editions can be seen as even more restricted by censorship and/or the need for well-equipped libraries.

as a further form of user interaction, or through subsequent republication. Corrections or improvements are also easily possible.[30] Users can (more) easily process the edition data for new questions that cannot be answered by simply reading and develop and apply new digital methods in general. At the same time, a digital edition can be linked to publications from other research projects in the sense of *linked open data.* In general, digital editions are characterized by a pronounced processualism in their production as well as in their use, and thus their product is in principle fluid in contrast to static print.

Digital editions offer another decisive advantage over printed editions: they are characterized by greater verifiability because they provide *raw* data from the editing process (such as transcriptions or collations) and can document the transmission and editorial decisions without space restrictions. Editors should make use of this possibility in the interests of transparency and for the sake of better editions.

4. Conclusion: What is still missing?

4.1 Education

To exhaust the potential of a digital edition, not only technically skilled editors are required, but above all a digitally literate user. The latter uses the digital edition (and the tools offered) with understanding and, depending on their profile or purpose of use, need not limit themselves to the presentation provided by the editors. The user can examine the data in the edition themselves for their own research questions using tools other than those offered.[31]

4.2 Tools

Digital editions in theological disciplines often deal with source materials or traditional contexts and are available in languages or language levels for which tools such as text recognition or natural language processing are not yet available or not available in sufficient quality.

30 Assmann & Sahle (2008) emphasize the new role of reviews of digital editions, since criticism and suggestions for improvement made in reviews, in contrast to printed editions, can actually lead to improvements because, unlike in printed editions, there is more than the theoretical possibility of a “new edition.”

31 Training and further education are particularly important for theological research, not least because the requirements of research funding for edition projects often exceed the capabilities of those involved and “the digital” should not and cannot simply be delegated entirely to IT or DH departments.

4.3 Cooperation and Standardization

Since theology has a long tradition as a science (orig. “Wissenschaft”), many relevant sources are already available in (printed) editions. The availability of printed resources is one reason why what is now being edited digitally often concerns either new discoveries or subjects that have not yet been well researched at the margins of the disciplines. There seems to be greater potential for innovation in these disciplines. In addition, third-party funding can motivate each edition project to publish its edition in its own web publication, which is then often a solitaire or presents itself as the proverbial silo whose contents cannot be examined in connection with other editions, not least because only in comparatively few cases is the data made accessible in addition to the presentation. In general, there is still a lack of overarching research options or at least a uniform editorial approach and (minimum) standards for the user interface, i. e., the large edition series known from book printing with their uniform editorial standards and layouts are missing.[32] Sometimes even the standard TEI makes interoperability between edition projects more difficult because, as a historically evolved and, above all, highly diverse standard, TEI offers several ways of labeling a phenomenon to be edited.

4.4 Credibility and Recognition

Digital editions in the field of theology have so far often been published as *hybrid* editions, i. e., in addition to the digital presentation, there is also a printed version of the edition. A potential reason for this hybrid publication strategy is not only the concern about the long-term availability and addressability of the digital edition, which is difficult to guarantee, but also the fact that digital publications are often not yet accepted by scholars where the digital editions are cited (and not their printed derivative) or generally recognized as scholarly publications.

32 It remains to be seen whether initiatives such as the *National Research Data Infrastructure* (https://www.nfdi.de) and, in particular, *Text+* (https://www.text-plus.org) can provide a remedy.

References

Apollon, D., Belisle, C., & Régnier, Ph. (Eds.). (2014). *Digital critical editions. Topics in the digital humanities.* Urbana/Chicago/Springfield: University of Illinois Press.

Arnold, E., & Döhnert, A. (2024). Editionspraxis 2021. Eine Disputation. In K. Bracht, A. Jouravel & J. Sieber (Eds.), *Methodius of Olympus. De lepra. Interdisciplinary Approaches* (pp. 145–156). Berlin/Boston: De Gruyter [= *Texte und Untersuchungen zur Geschichte der altchristlichen Literatur,* 189]. DOI: https://doi.org/10.1515/9783111350790-012 (Accessed: 18 June 2024).

Assmann, B., & Sahle, P. (2008). *Digital ist besser. Die Monumenta Germaniae Historica mit den dMGH auf dem Weg in die Zukunft – eine Momentaufnahme.* Norderstedt: Books on Demand GmbH [= *Schriften des Instituts für Dokumentologie und Editorik,* 1]. URL: http://kups.ub.uni-koeln.de/2317 (Accessed: 18 June 2024).

Barabucci, G., Spadini, E., & Turska, M. (2017). Data vs. Presentation. What is the core of a Scholarly Digital Edition? In P. Boot, A. Cappellotto, W. Dillen, F. Fischer, A. Kelly, A. Mertgens, A.-M. Sichani, E. Spadini, & D. Van Hulle (Eds.), *Advances in Digital Scholarly Editing. Papers presented at the DiXiT conferences in The Hague, Cologne, and Antwerp* (pp. 37–46). Leiden: Sidestone Press.

Bleier, R., Bürgermeister, M., Klug, H. W., Neuber, F., & Schneider, G. (Eds.). (2018). *Digital Scholarly Editions as Interfaces.* Norderstedt: Books on Demand, 2018 [= *Schriften des Instituts für Dokumentologie und Editorik* 12]. URL: https://kups.ub.uni-koeln.de/9085 (Accessed: 18 June 2024).

Boot, P., Capellotto, A., Dillen, W., Fischer, Franz., Kelly, A., Mertgens, A., Sichani, A.-M., Spadini, E., & Van Hulle, D. (Eds.). (2017). *Advances in Digital Scholarly Editing. Papers presented at the DiXiT conferences in The Hague, Cologne, and Antwerp.* Leiden: Sidestone Press.

Brandenburg, Y. (2020). A Review of confessio.ie, or Practical Thoughts on Digital Editing in Classics, *RIDE,* 13, 1–28. DOI: https://doi.org/10.18716/RIDE.A.13.5 (Accessed: 18 June 2024).

Cummings, J. (2018). A world of difference. Myths and misconceptions about the TEI, *Digital Scholarship in the Humanities,* 34(1), 58–79. DOI: https://doi.org/10.1093/llc/fqy071 (Accessed: 18 June 2024).

Driscoll, M. J., & Pierazzo, E. (Eds.). (2016). *Digital Scholarly Editing. Theories and Practices.* Cambridge: Open Book Publishers. URL: http://www.openbookpublishers.com/product/483 (Accessed: 18 June 2024).

Kuczera, A. (2022). TEI Beyond XML. Digital Scholarly Editions as Provenance Knowledge Graphs. In T. L. Andrews, F. Diehr, Th. Efer, id. & J. J. van Zundert (Eds.), *Graph Technologies in the Humanities. Proceedings 2020* (pp. 101–123). Wien: CEUR Workshop Proceedings. URL: https://ceur-ws.org/Vol-3110/preface.pdf (Accessed: 18 June 2024).

Meins, F. (2016). Digitale Editionen in den Altertumswissenschaften?, *Digital Classics Online,* 2, 49–57. DOI: https://doi.org/10.11588/dco.2016.1.24491 (Accessed: 18 June 2024).

Neill, I., & Kuczera, A. (2019). The Codex. An Atlas of Relations. In A. Kuczera, Th. Wübbena & Th. Kollatz (Eds.), *Die Modellierung des Zweifels. Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten* (no pag.). Wolfenbüttel: Herzog August Bibliothek [= *Zeitschrift für digitale Geisteswissenschaften. Sonderbände.* 4]. DOI: https://doi.org/10.17175/sb004_008 (Accessed: 18 June 2024).

Pierazzo, E. (2015). *Digital scholarly editing. Theories, models and methods.* Farnham/Burlinton: Ashgate.

Sahle, P. (2013a). *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels.* Teil 1. Das typografische Erbe. [Preprint-Fassung]. Norderstedt: Books on Demand GmbH [= *Schriften des Instituts für Dokumentologie und Editorik,* 7]. URL: https://kups.ub.uni-koeln.de/5011 (Accessed: 18 June 2024).

Id. (2013b). *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels.* Teil 2. Befunde, Theorie und Methodik. [Preprint-Fassung]. Norderstedt: Books on Demand GmbH [= *Schriften des Instituts für Dokumentologie und Editorik,* 8]. URL: https://kups.ub.uni-koeln.de/5012 (Accessed: 18 June 2024).

Id. (2013c). *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels.* Teil 3. Textbegriffe und Recodierung. [Preprint-Fassung]. Norderstedt: Books on Demand GmbH [= *Schriften des Instituts für Dokumentologie und Editorik,* 9]. URL: https://kups.ub.uni-koeln.de/5013 (Accessed: 18 June 2024).

Id. (2014). Kriterienkatalog für die Besprechung digitaler Editionen (Version 1.1). In *Institut für Dokumentologie und Editorik.* URL: https://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1 (Accessed: 18 June 2024).

Id. (2016). What is a Scholarly Digital Edition? In M.J. Driscoll & E. Pierazzo (Eds.), *Digital Scholarly Editing. Theories and Practices* (pp. 19–40). Cambridge: Open Book Publishers. DOI: https://doi.org/10.11647/OBP.0095.02 (Accessed: 18 June 2024).

Schnöpf, M. (2014). Codex Sinaiticus, *RIDE,* 1, 1–21. DOI: https://doi.org/10.18716/RIDE.A.1.2 (Accessed: 18 June 2024).

Stockhausen, A.v. (2020). Die Modellierung kritischer Editionen im digitalen Zeitalter, *Zeitschrift für Antikes Christentum,* 24(1), 123–160. DOI: https://doi.org/10.1515/zac-2020-0019 (Accessed: 18 June 2024).

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, Ph.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, Ch.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., C't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, Ph., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, Th., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship, *Scientific Data,* 3(1), 1–9. DOI: https://doi.org/10.1038/sdata.2016.18 (Accessed: 18 June 2024).

V. DISSEMINATION

Scientific Communication and Community Building

Ulrike Wuttke

https://orcid.org/0000-0002-8217-4025

Abstract This chapter provides an overview of the manifestations and specifics of scientific communication and community building in the Digital Humanities. After a brief historical introduction, the chapter presents important definitions, actors, and focal points, as well as tools and initiatives. Despite innovation, the chapter shows how the opportunities of the digital paradigm in science communication are not yet fully being utilized and what can be done in the future to bring about change. Moreover, potentials and expertise of the Digital Humanities in this field, e.g., about the connection between ethical aspects and technological issues, are outlined. The chapter also shows how closely scientific communication and community building are related, particularly through social media.*

Keywords Scholarly Communication, Digital Publishing, Community Building, Digital Humanities

DISSEMINATION – INTRODUCTION TO THE SECTION

Communication is an essential part of science: researchers communicate both with each other and with the wider public or particular communities. While the intention and manifestation of the various acts of communication may differ depending on the stage of the research process, target group(s), and degree of formalization, the common threads at the heart of the scholarly communication process are questions of dissemination and access to scholarly information (De Silva & Vance 2017, 17 f.) as well as the emergence of scholarship.

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

Ulrike Wuttke: Scientific Communication and Community Building, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 347–378. DOI: https://doi.org/10.11588/heibooks.1521.c21959

1. Standing on the Shoulders of Giants

In the 17th century, Isaac Newton claimed he had only succeeded in progressing further (i.e., achieving scientific progress) because he was standing on the shoulders of giants (i.e., because he could build on the works of those who had gone before).[1] The contemporary understanding of science is based on this fundamental principle.[2] However, according to Hagenhoff et al. (2007), the characteristics and requirements on the mechanisms of proof and access have changed fundamentally since Newton:

1. through the digital transformation of science (communication) and the associated change in the scientific record *(changing scholarly record)* as well as
2. the increasing openness of scholarly communication and the academic system, i.e. the greater role of open principles in communication and research processes.

2. Dissemination within the Humanities

The epistemic classification of the term *dissemination* in the knowledge organization of the Digital Humanities can be made in the context of the so-called "scholarly primitives" or "methodological commons" (Unsworth 2000; cf. Van der Weel & Praal 2020; Borek et al. 2021). The concept has been indicated as a component of the "digital scientific process and for the shaping of knowledge" (Borek et al. 2021, 322) as an upper category in the *Taxonomy of Digital Research Activities in the Humanities* (TaDiRAH[3]) and refined by subcategories *(narrower concepts)* such as *collaborating, commenting, communicating, crowdsourcing, publishing, sharing,* and *teaching*. The category *dissemination* is broadly defined in TaDiRAH with a focus on (co)sharing:

> disseminating refers to the activity of making objects of inquiry, results of research, or software and services available to fellow researchers or the wider public in a variety of more or less formal ways.[4]

This chapter outlines the main features of scientific communication and community building with a focus on the Digital Humanities to introduce the topic of dissemination

1 S. De Silva & Vance 2017, 101 f. The phrase is often falsely attributed to Isaac Newton. It can actually be traced back to Bernhard of Chartres (12th Century) (Chen 2003, 135–166).
2 The functions of *scientific communication* have traditionally been analysed in terms of four key categories: *"registration, awareness, certification, and archive functions"* (Rosendaal & Geurts 1999, 14).
3 See https://tadirah.info (Accessed: 19 June 2024).
4 See https://vocabs.dariah.eu/tadirah/en/page/disseminating (Accessed: 19 June 2024).

in this compendium and to discuss certain aspects of the topic of dissemination. The focus is on an outline of historical developments, current key topics, and prospects for scientific communication and community building, as well as an examination of the conceptual connection between the two terms. Some of the aspects listed here are dealt with in greater depth in the following chapters of this compendium.

SCIENTIFIC COMMUNICATION

This section offers a definitional approach to the concept of scientific communication, including a brief social, science policy, and historical context, as well as an introduction to selected key topics.[5]

1. Localization by Definition

The terms scholarly or *scientific communication* (orig. "Wissenschaftskommunikation"), *science communication* (orig. "Wissenskommunikation"), and knowledge transfer can be distinguished from one another, even if there are overlaps and the terminology is currently in flux (Wissenschaftsrat 2016; Schuldt-Baumgart 2022).[6] The term *science communication* focuses narrowly on the impact of research in society and *knowledge transfer* on the systematic and targeted transfer of knowledge to the economy and society (Schuldt-Baumgart 2022). In contrast, the target groups, and goals of scientific communication, such as informing, sensitizing, inspiring, strengthening the reputation or legitimizing science, are more heterogeneous (ibid.).

In a broader sense, the term *science communication* refers to different forms of communication by and about science, the common denominator being scientific processes, methods, practices, and publications, e.g., articles or monographs, but also preliminary stages, data, and other communication formats aimed at different target groups. Traditionally, there is a distinction between internal and external science communication according to the sender-receiver principle (Fig. 1).[7] Internal science communication is aimed at internal scientific target groups (specialist public) and

5 Scientific communication can manifest itself in various ways. See, for example, COAR (https://www.coar-repositories.org) and DORA (https://sfdora.org). Both addresses were accessed on 19 June 2024. The chapters by C. Anderson (digital forms of publication) and J. Apel (research data) in this volume provide more information on these aspects.

6 In English publications, there is not always a clear distinction between *science communication, scholarly communication,* and *scientific communication.*

7 There are also other traditional approaches to the systematization of science communication, such as, overall system, institution, individual (macro, meso, and micro levels) (Dernbach et al. 2012).

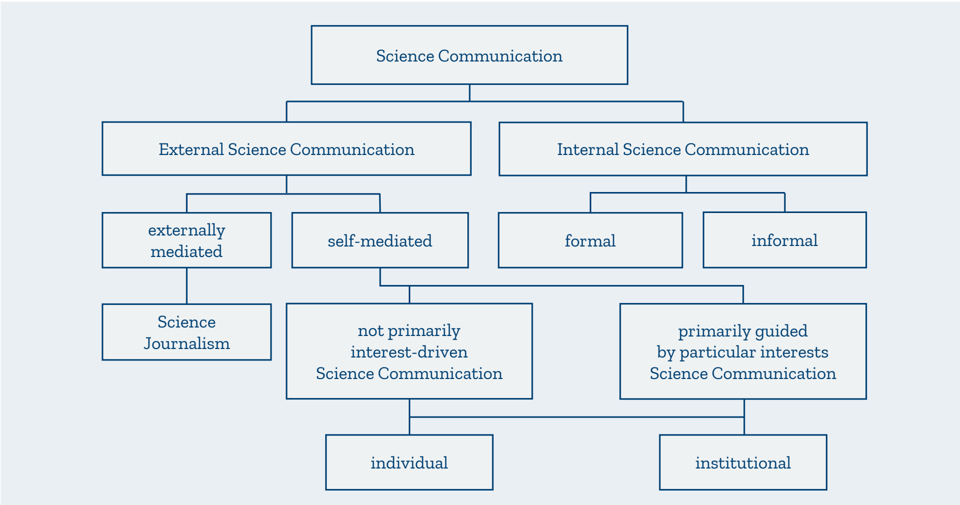


Fig. 1 Traditional systematization of science communication.

external science communication at target groups outside science (acatech 2017, 20–21; cf. Pasternack 2022, 42). However, the excessive emphasis on or equation of science communication with external science communication in current science and socio-political statements and policy papers, such as those of the Federal Ministry of Education and Research (BMBF 2019, 2) or the German Council of Science and Humanities (2021, 7), can be viewed critically due to the independent functions of internal and external science communication.

The distinction between formal and informal internal science communication (publication versus personal contact) is particularly important for community building (see below). The various activities can be modeled in an increasingly digitized and network-based publication cycle with the steps of writing, reviewing, publishing, storing and making accessible (library, repository), receiving, citing, annotating (etc.) (Umlauf & Gradmann 2014). Moreover, various dimensions of science communication can be considered, such as content, target group, style, format, motivation, and own role (Seltmann 2023, 2). A recent systematization approach proposes a stronger focus on the senders of science communication and based on this, distinguishes between one level of internal science communication *(science-to-science)* and three levels of external science communication *(science-to-public, public-to-science,* and *public-to-public)* (see Fig. 2) (Frick et al. 2021; Seltmann 2023, 2).

As a result of digitalization, the systematization of science communication, together with the structures of the science communication system, is undergoing major changes (acatech 2017, 20–21). Phenomena, such as open access, internet publication platforms, and social media are increasingly leading to “a convergence of different

Receiver / Sender	Science	Public
Science	Science-to-Science	Science-to-Public
Public	Public-to-Science	Public-to-Public

Fig. 2 Four-part matrix of science communication.

forms of science communication" (acatech 2017, 21; cf. Weitze & Heckl 2016, 191). For this reason, the following section avoids sharp distinctions and instead focuses on selected key topics of science communication in the (Digital) Humanities with a focus on forms of science communication in which scientists themselves are directly involved as communicators.[8]

2. Science and Sociopolitical Positioning

In the context of science and socio-political discussions, external science communication has recently received renewed importance for public discourse and promoted as a task in the science system (Baumgärtner et al. 2021). Regarding this politically desired cultural shift (cf. BMBF 2019), a statement by humanities associations critically discussed the danger that public scientific discourse would only promote topics relevant to the public (Baumgärtner et al. 2021). Instead of an undifferentiated call for *more* science communication, the statement advocated for a differentiated, fundamental reflection on the goals, motives, expectations, and contexts of science communication (ibid., on the goals of science communication, cf. Ziegler & Fischer 2020). The critical discussion of the evaluation of research relevance by Fecher (2022) is similar. Proposed solutions include moving away from the so-called watering can principle and superficial approaches, such as camera training or social media workshops (ibid.). Instead, there should be a shift towards specific training for context-specific, target group-oriented, and problem-conscious science communication aimed at sub-groups, as well as the creation of fruitful framework conditions (keyword: serendipity) (ibid., cf. Frick & Seltmann 2023).

Technological developments, such as the latest generation of AI-supported text, image, or sound generation tools or 3D/VR technologies create new possibilities for science communication, yet they also must be critically evaluated. The discussion

8 Doing so largely excludes citizen-science based *public-to-public* and *public-to-science* forms of science communication (Hecker et al. 2018), as well as forms of science PR (i.e., external, institutional science communication primarily guided by interests) or science journalism (i.e., external, externally mediated science communication) (Wissenschaftsrat 2021).

about their potential and dangers is not only about their practical applications, but also about aspects of communication with society via such technologies (Schröder 2023). It is conceivable that scholars engaged in Digital Humanities could assume an influential position in this context.

3. Historical Aspects

Since the 17th century, the academic publication system has developed into a standardized and familiar form. Academic journals are crucial for formalized, internal academic communication (De Silva & Vance 2017, 17–24). A discipline-specific feature of the publication culture in the humanities is the formative role of books (monographs, edited volumes, etc.), hence the demand for these avenues of research to be considered alongside journal articles in science policy discourses, e. g., on open access (Söllner 2017; Winters 2020; Toledo 2020).

The commercialization of the publication market in the 21st century has led to a shift in focus from the scholarly societies that originally dominated the market to commercial players in the form of academic publishers (as monopolies or oligopolies), while at the same time the volume of publications has risen sharply since the Second World War (De Silva & Vance 2017, 17–24). Even if commercialization is less advanced in the humanities than in other disciplines (Larivière et al. 2015), Unsworth (2003) also painted a bleak picture for the humanities (esp. with regard to promotion and tenure procedures): “here [there] seems to be general agreement that the system of scholarly communication is not working – that it is broken, or breaking.” He critically suggested that a better scholarly communication system should be based on the premises of appropriateness of the form of communication of research results and user comfort. Ultimately, the evaluation of research should be based on content quality criteria and impact, not on quantity or form (analog or digital):

> In a better world, high-quality, peer-reviewed information would be freely available soon after its creation; it would be digital by default, but optionally available in print for a price; it would be easy to find, and it would be available long after its creation, at a stable address, in a stable form (Unsworth 2003).

The following comments on key issues in scholarly communication in the (Digital) Humanities underline the continuing relevance of Unsworth’s criticism, which is now twenty years old (cf. Heise 2018; Edmond 2020; Eve & Gray 2020; AG Digitales Publizieren 2021). The order of the key topics does not express any empirically substantiated value statement.

4. Key Issues

4.1 Changing Scholarly Record

The publishing culture in the humanities is strongly characterized by the print paradigm and reservations about digital publishing, such as doubts about long-term archiving and availability or quality assurance, have a negative impact on the acceptance of open access publications (Winters 2020, 345–347). Fluid and hybrid digital forms of academic publication that go beyond PDF are slowly gaining acceptance, partly because of the advantages associated with the change of media, such as "machine readability, multimedia, modifiability, easy copying, networkability, etc." (AG Digitales Publizieren 2021, para. 4; Winters 2020). Digital edition philology with the TEI guidelines has been established as the standard for digital editions, while remaining subject to dynamic development processes and subsequent requirements (see the chapter by A. v. Stockhausen in this volume, p. 335). One could first mention the possibilities of podcasts (from *iPod* and *broadcast*) as an example of more recent trends in the medial expansion of the publication spectrum,[9] which not only offer interesting perspectives for addressing different sub-populations and popularizing academic topics but can also influence the research process itself (Howard-Sukhil et al. 2021). In addition, there are other communication formats or research products that can be published along the open research cycle in accordance with the FAIR principles (see the chapter by J. Apel in this volume): Pre-registrations via *OSF Registries*[10], Preprints via *Preprints.org*[11] (general), *BodoArXiv*[12] (medieval) (cf. Dang 2017), or Open Peer Review Reports (see below). The possibility of linking publications with research data, software or code, or publishing them independently is also important (see the chapters by J. Apel and U. Henny-Krahmer in this volume).

4.2 Forms of Authorship

Throughout the digital transformation, the forms of publication and their mediality have changed as well as the forms of authorship "when knowledge processes are more collaborative and understood as dynamic and mapped transparently" (AG Digitales Publizieren 2021, para. 18). Collaborative, social writing processes and the verification of different roles are promoted by appropriate tools (e.g., wikis) or

9 Like most social media platforms, the podcast landscape is rapidly changing. Cf. RaDiHum20 (https://radihum20.de), Humanista (https://humanistathepodcast.com), and Price Lab Podcast (https://pricelab.sas.upenn.edu/podcast). All addresses were accessed on 19 June 2024.

10 See https://osf.io/registries (Accessed: 19 June 2024).

11 See https://www.preprints.org (Accessed: 19 June 2024).

12 See https://bodoarxiv.wordpress.com (Accessed: 19 June 2024).

taxonomies (e.g. *CRedditT*[13]). One desideratum is better support through the development of adapted reference systems and citation methods (AG Digitales Publizieren 2021, para. 18–22; Ernst 2015).

4.3 Peer Review

Mechanisms to ensure the critical, neutral evaluation of scientific claims and sources are crucial in the scientific system, with efforts to increase quality and ensure the credibility of science (as opposed to fake science) being important drivers (De Silva & Vance 2017, 73–99). These tasks are performed by peer review (ibid., 74). Formalized peer review procedures are used for various publication products (e.g., *pre-publication* & *post-publication peer review*) and other research contexts (e.g., in the evaluation of research proposals and performance and other competitive selection procedures) (ibid., 73–99).

Although the organization and usefulness of peer review procedures are subject to critical debate, both in general (ibid., 74) and subject-specific terms (AG Digitales Publizieren 2021, paras. 45–55), formal review procedures are becoming the norm in the humanities. In an interdisciplinary field, such as the Digital Humanities, the bias against interdisciplinary research poses a challenge (ibid., 74); there are also differences in each discipline about the acceptance of different peer review procedures, e.g., closed peer review *(single blind* and *double blind)* and open peer review *(pre-* and *post-publication)* (ibid., 81). To be more transparent, open peer review procedures are gaining popularity (e.g., in the *Zeitschrift für digitale Geisteswissenschaften – ZfdG*)[14] as are aspects of the separation between content-related, formal, and technical review procedures (ibid., 82–86; Ross-Hellauer 2017; AG Digitales Publizieren 2021, paras. 54–57; Burghardt et al. 2022). In this context, preprint servers have become increasingly relevant in the humanities (Kleineberg & Kaden 2017).

4.4 The Concept of Publication

In humanities research, there is currently a controversial debate about the extent to which the concept of publication is linked to peer review or other formal quality assurance procedures, i.e., the extent to which they are considered a prerequisite for assessment as a publication (Edmond & Romary 2020).[15] While the use of formalized review procedures can lead to more confidence in the quality of digital publications (AG Digitales Publizieren 2021, para. 57), one must continue to discuss the praxis of

13 See https://credit.niso.org (Accessed: 19 June 2024).
14 See https://www.zfdg.de (Accessed: 19 June 2024).
15 See also the categorization of open access models, expressed using the colors grey, green, and gold.

recognition or evaluation criteria regarding the recognition of a greater diversity of media manifestations of science communication, as expressed by DORA.[16]

4.5 Open Access to Scientific Information

Since the 1970s, the commercialization of the journal market and, at times, significant increases in subscription costs, especially in the STEM field (the so-called journal crisis), have posed enormous challenges for academic libraries and other actors in the public information infrastructure, which are among the decisive factors for the emergence of the open access movement and the pursuit of open access to information (De Silva & Vance 2017, 17–40).[17] In addition, mega-journals, such as *PLOS One* or *Scientific Reports,* both more active in STEM, emerged as a new business model (Davis 2017). The focus of the largest scientific publishers has shifted from the sale of publishing products to data analytics and the sale of usage data. The risk of the development of proprietary workbenches or workflows by these actors (Bosman & Kramer 2018) for the science system has recently been increasingly recognized as problematic, so far without comprehensive consequences (Couldry & Mejias 2019; AWBI 2021; Kunz 2022).

Even if accepting open access paradigms presents the humanities with challenges specific to the academic discipline (cf. Heise 2018), they are now largely established (Söllner 2017; Kleineberg & Kaden 2017; DARIAH-EU 2018; Heise 2018; Wuttke & Gebert 2021; AG Digitales Publizieren 2021, paras. 79–111). Open access publication models and initiatives, such as the *scholarly led Open Library of Humanities* (OLH) or the *non-profit, academy owned* open access initiative *AmeliCA* from the Global South (Becerril-García 2019), are valuable additions to the publication spectrum, and most humanities journals listed in the DOAJ *(Directory of Open Access Journals)*[18] do not charge APC. An important goal of a scientific practice characterized by principles of openness is to make the research process more transparent, e.g., through the publication of work-in-progress, interim results, and products (in contrast to the traditional focus on research results), even involving external parties in the generation of research ideas *(open innovation),* whereby the opening up of academic communication through alternative, attractive, and comprehensible forms that are accessible to a wider audience plays an important role (Niemann et al. 2017; Wuttke & Gebert 2021, 436). For this reason, the development of alternative criteria and recognition

16 DORA stands for "Declaration of Research Assessment." Cf. https://sfdora.org (Accessed: 19 June 2024).

17 With reference to the discussion about the term "publication" (see above), the term "information" was deliberately chosen to symbolize greater diversity.

18 See https://doaj.org (Accessed: 19 June 2024). 192 of the 215 journals registered in the DOAJ do not charge APC (Accessed: April 2023).

mechanisms for science communication, such as strengthening the transparency or credibility of science, is necessary (see below). Conversely, this development also includes increasing the visibility and possibility of access to quality certified scientific publications in open access, so that interested parties are not blocked by a paywall. The importance of these goals has now been declared a global priority by UNESCO.[19]

4.6 Bibliodiversity

Bibliodiversity stands for efforts to maintain greater diversity regarding the forms of expression of academic communication and, in the context of the humanities, stands, among other things, for the strengthening of multilingualism in view of the increasing dominance of the English language (Balula et al. 2021; Balula & Leão 2021). The shift towards a monolingual, English-dominated publication and academic landscape driven by commercial publishers, but also by science policy and science-internal actors, jeopardizes the engagement with different cultures and peculiarities that often takes place in humanities and social science contexts as well as the inclusion of the broader public (Shi 2023).[20] This monolingual dominance also contributes to the divide between the global North and the global South and is thus at odds with the 17 UN Sustainable Development Goals.[21] The *Jussieu Call for Open Science and Bibliodiversity,* for example, calls for the strengthening of bibliodiversity.[22]

4.7 Quantification of Science (Impact Factors & Co.)

In view of the increasing amount of information, there is a desire for objective, quantitative mechanisms that can replace time-consuming qualitative evaluation procedures (De Silva and Vance 2017, 101 f.). On the level of journal articles, the measurement of citation-based impact is based on the presumed correlation between the number of citations and the *impact* of a publication, however, this result should be viewed critically due to fundamental problems (ibid.). The same applies to the *Journal Impact Factor* (JIF), which was originally developed to assess the quality of journals and is now often incorrectly used to assess individual research performance of research articles (ibid., 104–108). Moreover, prevalent citation indexes can often only

19 UNESCO recommendation on Open Science (2021): https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en (Accessed: 19 June 2024).

20 Interesting, yet beyond the scope of this chapter, is the critical discussion of the alternative positive view of Latin as a language of science and the negative view of national-language science and the resulting welcome for the establishment of English as a language of science (Voigt 2012, 9).

21 See https://sdgs.un.org/goals (Accessed: 19 June 2024).

22 See https://jussieucall.org/jussieu-call (Accessed: 19 June 2024).

be used fully functionally for a fee (e.g., Web of Science, Scopus), which is why open citation indexes are being parallelly developed (Peroni & Shotton 2020).[23]

The aim of the Metrics-Literacies Project (Maggio et al. 2022)[24] is to curb the careless use of these methods in light of the critique of bibliometric processes, such as the JIF or the h-index, as well as of the use of alternative indicators (altmetrics), for assessing the quality of individual publications or the research performance of individual persons or research groups, the acquisition of a basic understanding of the most common methods, and their justified points of criticism and disciplinary differences in the context of academic research. In addition, these are internal scientific factors that contribute little to the broader impact of research (De Silva and Vance 2017, 109–112; Wróblewska 2021). Regarding the evaluation of science communication beyond bibliometric methods, there is a rich spectrum of methods only slowly arriving in practice (Niemann et al. 2023).

4.8 Social Media

The increasing role of social media is understood as a key factor in blurring the boundary between internal and external science communication (acatech 2017, 11). Social media includes:

> individual formats such as blogs and podcasts, which are usually operated by one person or organization, as well as collective formats such as social network sites (SNS, such as Facebook), microblogging services (such as Twitter), video and photo platforms (such as YouTube or Instagram) and wikis (such as Wikipedia), in each of which a large number of networked users participate within a single contribution (acatech 2017, 11).[25]

In the decentralized Web 2.0, the so-called participatory web, anyone can be a sender and receiver, for which the term *prosumer,* an artificial word made up of *consumer* and *producer,* was coined. The advantages and disadvantages of this development are directly relevant to science communication on the one hand yet also go beyond this narrow context (Peters 2023; cf. Voigt 2012, 9–10). Some advantages include the empowerment of individuals to generate extensive reach independently of gatekeepers like publishers, journals, etc., while the disadvantages include the collection and evaluation of data (e.g., through commercial data tracking, esp. big data), dependence

23 See http://opencitations.net/index (Accessed: 19 June 2024).
24 See https://www.scholcommlab.ca/research/metrics-literacies (Accessed: 19 June 2024).
25 There is possibly also a difference between social networks and content-sharing services (Peters 2023).

on non-transparent algorithms geared towards optimizing advertising, the risk of political manipulation, and quality assurance issues (Schöch 2016; Könneker 2020).

Anyone can participate in the main functions of social media, such as information dissemination (creating and publishing) and information transfer (networking, commenting, annotating, sharing, and subscribing), thereby becoming part of the diffusion of information (Peters 2023). In this respect, overlaps between science communication processes and social aspects of community building become apparent (Seltmann 2023, 3). However, the increasing commercialization of Web 2.0 reinforces existing tendencies towards the commercialization of scholarly communication and leads to conflicts under copyright or data protection law. This process is based on open protocols and accomplished by proprietary internet platforms and services, in particular through the so-called platform economy, including social media (Peters 2023, 691). As a result, non-profit publication and social media platforms and infrastructures are becoming increasingly important (AG Digitales Publizieren 2021, paras. 90f.). These include, for example, approaches such as ORCiD (digital author profile, PID),[26] (academic) social media (e.g., *Humanities Commons*),[27] academic blogs,[28] or tools of the so-called Fediverse[29] as alternatives to commercial social media platforms (Brembs et al. 2023).

By focusing on user-generated content, social media call into question traditional academic publication and communication practices, which can lead to reservations about their academic nature within the humanities establishment (König 2015; Geier & Gottschling 2019). It is not always easy to prove to what extent and with what effect social media activities by academics reach target groups inside and outside academia. However, a lack of formal recognition mechanisms and a lack of practical skills (Könneker 2020) can have a disparaging effect and lead to their inherent potential remaining underdeveloped (König 2015). One negative effect can be the currently low participation of German scientists with their specialist expertise in social media discourses and thus their lack of a counterpoint to populist figures (Könneker 2020). At the same time, science communication on such channels is suitable for promoting internal and direct public dialog and thus for stimulating a broad exchange on scientific projects, methods, and findings (Geier & Gottschling 2019, 284).

26 See https://orcid.org (Accessed: 19 June 2024).

27 See https://hcommons.org (Accessed: 19 June 2024).

28 Such as on *Hypotheses*, the blog channel for the humanities and social sciences.

29 The *Fediverse* is a network of federated social networks and other services based on the open communication protocol ActivityPub (Wikipedia 2023, König 2022).

4.9 The Role of Libraries

Libraries function as “service providers and laboratories of science communication” (Frick et al. 2021), particularly in the non-profit area of science communication.[30] For the humanities, this pairing is a continuation of a long tradition. As so-called memory institutions, the collection, preservation, and availability of information is one of the most fundamental tasks of libraries, whose range of tasks has shifted in the course of the digital transformation and the increasing importance of open science principles from safekeeping and preservation (including reference systems) to active support of the entire publication cycle, including publication processes (OJS, repositories, consulting, university publishers) (Neuroth 2017; AG Digitales Publizieren 2021, para. 15). These activities also include current developments such as the hosting of decentralized Mastodon servers in the Fediverse, e.g., by the Max Planck Digital Library[31] or the Staatsbibliothek zu Berlin.[32] The course of these developments is difficult to predict; however, the dedicated positioning as places of knowledge and its curation offers libraries a promising perspective for the future (Brembs & Siems 2023).

5. Interim Conclusion

On the one hand, the way the academic system is opening up along various parameters is leading to an increasing diversity of products, channels, and target groups. However, on the other hand, Digital Humanities scholars are clinging to traditional paradigms of academic communication despite a claim to innovation (Nyhan 2020; Sahle & Neuber 2022).

Scholarly communication is an intrinsic part of the Digital Humanities, and there are many corresponding activities that deal with topics of scholarly communication and digital publishing. Future theory development will show whether terms like *Public Humanities* will prevail (Burghardt 2020; Gundermann et al. 2021; Schwan & Thomson 2022; Seltmann 2023).

30 Hence why the development of competence in this area is becoming increasingly important (Frick & Seltmann 2024).

31 See https://social.mpdl.mpg.de (Accessed: 19 June 2024).

32 See https://openbiblio.social/about (Accessed: 19 June 2024).

COMMUNITY BUILDING

This subsection provides an approach to defining the term *Community Building,*[33] including historical developments, as well as its fundamental aspects (networks, resources and tools, training and education, and the promotion of collaborative projects and initiatives).

1. Definition

A comprehensive theoretical localization of the term *Community Building* has yet to be recorded in the Digital Humanities. The literature mainly deals with practical approaches, opportunities, and challenges (Busch et al. 2016; Prescott 2016; Fitzpatrick 2020).

The following analysis argues that at the heart of Community Building are social processes related to networking and science communication. Ultimately, given the tools of social media and Web 2.0, the relevant functions of Community Building are increasingly overlapping and lead to blurred boundaries between academic communication and community building. One example is *academia.edu,*[34] a scientific social network mainly used by scientists for two functions:

1. Science communication: promoting and making available their own publications (and other scientific achievements) (legally controversial due to possible copyright infringements associated with this); and
2. Community Building: forms of scientific networking.

However, the academic community is highly critical of *academia.edu* due to its opaque algorithms and access logic and the commercial business model based on them, as well as data protection and copyright issues (Schöch 2016; Fitzpatrick 2020, 351–353).

Community Building has both individual and collective elements. Understood from an interdisciplinary perspective, however, a perspective transcending the individual is more relevant from a sociology of science perspective. The focus on collective aspects, i.e., associations of like-minded scholars with different degrees of institutionalization, Community Building in the Digital Humanities is then based not only on meta-level oscillation between the humanities and computer science, but

33 The German equivalents of the term are *Gemeinschaftsbildung* and *Gemeinschaftsaufbau.* Usually, however, the english term is also used here.

34 See https://www.academia.edu (Accessed: 19 June 2024).

also between different disciplines within the Digital Humanities (Sahle 2015; Benatti et al. 2021, paras. 14 f.). Due to the intrinsic interdisciplinarity (Klein 2015), digital scholars are faced with the challenge of networking beyond their respective narrower scholarly societies. The exchange across traditional disciplinary or organizational boundaries is crucial for strengthening the field of the Digital Humanities, particularly for the exploration of common research interests (Wuttke 2022, para. 53).

The explorations in this subsection are based on the following working definition of Community Building in the context of Digital Humanities:

> The term *Community Building* refers to activities and efforts to create suitable conditions for the development of an open and inclusive community of researchers, practitioners, students, and other interested parties within the Digital Humanities. Community Building aims to promote the exchange of ideas, knowledge, and resources. By building a strong community, the Digital Humanities can grow as a field, advance knowledge, and promote collaboration across disciplinary and geographic boundaries.

Four fundamental aspects of Community Building are explained in more detail below:

1. Community Building through networking,
2. resources tools and infrastructures,
3. training and education as a Community Building task,
4. collaboration, and Community Building projects.

2. Community Building Through Networking

The first aspect of Community Building concerns the creation of networks, platforms, and opportunities to share projects, ideas, and challenges and to learn, communicate and collaborate with each other. One aim is to raise awareness of the added value of digital research methods and to share good practices, projects, and infrastructures (Busch et al. 2016, 279). In networking activities, one needs to strengthen thematic dialog so that communication is not one-sided and the network "is not used purely as an information and marketing platform or for a one-sided information interest," but used to explore potentials like identifying needs, building expertise, interdisciplinary networking, and creating reach (external impact) (ibid., 281 f.).

In the history of science, the exchange of communication accelerated by the journals established in the 17th century is seen as initiating the development of the *scientific community* and the differentiation of scholarly societies (disciplines) (Voigt 2012, 13). However, access to this community was difficult for scientists outside an elite group (ibid.). Despite the justified criticism of the *invisible college* (ibid.), *scholarly* *societies* emerged as part of this process since the early modern period and have since become less elitist given their emphasis on Community Building.

Scholarly societies are formal, self-governing scientific organizations. They consist of individuals with the same values, with the historical aim of promoting scholarly communication between their members and the wider intellectual world, e.g., through newsletters, meetings, journals, and conferences (Fitzpatrick 2020, 353). They also offer opportunities for Community Building through conference participation, networking, and collaboration, which are highly valued by members (Winters 2020, 343). Examples of scholarly societies in the Digital Humanities include the DHd – *Association for Digital Humanities in German Speaking Areas* ("DHd-Verband"),[35] the *European Association for Digital Humanities* (EADH),[36] and the international umbrella organization *Alliance of Digital Humanities Associations* (ADHO)[37] (Prescott 2016).

Other examples of formalized associations include digital branches, working groups, initiatives within individual scholarly societies,[38] thematically relevant (inter)national research infrastructures (see below), and regional or local research associations such as *mainzed*[39] (located in Mainz), the Heidelberg *THEOLAB*[40], or Digital Humanities centers[41], such as *the Trier Center for Digital Humanities*[42]. Informal structures of low-threshold networking and support, such as DH groups, brown bag lunches, or ad hoc working groups, also have important functions as nuclei or multipliers (Burghardt & Wolff 2015; Roeder et al. 2019; Wuttke 2022, para. 53). The forms of networking activities and tools in the Digital Humanities and the specific challenges involved in establishing them are ultimately as diverse as their different target groups and framework conditions: Networking takes place through conferences and workshops, among other things, as well as increasingly through online forums, social media, and other digital communication channels (Estill et al. 2022; see below).

3. Resources, Tools, and Infrastructure

The provision of resources, tools, and infrastructure that meet the needs of the community also need to be considered under the aspect of Community Building, including but not limited to, the development of open-source software, the provision of

35 See https://dig-hum.de (Accessed: 19 June 2024).
36 See http://eadh.org (Accessed: 19 June 2024).
37 See https://adho.org (Accessed: 19 June 2024).
38 See https://dig-hum.de/initiativen-den-geisteswissenschaftlichen-fachcommunities (Accessed: 19 June 2024).
39 See http://mainzed.org (Accessed: 19 June 2024).
40 See https://theolab.hypotheses.org (Accessed: 19 June 2024).
41 See https://dhcenternet.org (Accessed: 19 June 2024).
42 See https://tcdh.uni-trier.de/de (Accessed: 19 June 2024).

databases, digital collections, and other digital resources. Sustaining these aspects is ultimately based on the community, i.e., on their provision as digital commons for the most open use possible (Dulong De Rosnay & Stalder 2020). Only their constant transfer to other contexts and updating through further development, scientific communication, or their use in teaching leads to true sustainability through appropriation by different communities as opposed to static preservation through "freezing" (Fenlon et al. 2023). By sharing access to tools and resources, community members can work more efficiently and learn from each other.

For the role of infrastructures in the Digital Humanities, such as CLARIN *(Common Language Resources and Technology Infrastructure)*[43], DARIAH *(Digital Research Infrastructure for the Arts and Humanities)*[44] or RESILIENCE *(REligious Studies Infrastructure. TooLs, Innovation, Experts, conNections and Centres in Europe),*[45] reference should also be made to the ESFRI roadmap at European level.[46] There are essential elements of non-commercial community spaces and services, including research infrastructures, academic societies, and academic libraries.

For a long time, exchanges between researchers were limited to internal forms of communication, i.e., informal forms such as letters or personal exchanges, the landscape was formalized in the 17th century with the founding of the first scientific journals. For a long time, informal communication was dominated by traditional methods such as circular letters, meetings, journals, and conferences. The scene changed drastically with the introduction of the internet, which brought forth applications and forms suitable for a wide range of people to exchange and disseminate informal news quickly and cost-effectively, while also expanding the tools of Community Building to include a digital spectrum, such as mailing lists (O'Donnell 2020). The *scientific community* became more tangible for individual scientists through the new technological possibilities of exchanging information with remote and scattered communication partners (Voigt 2012, 14).

Excursus: Academic Mailing Lists – From Science Communication to Community Building

This excursus outlines the changing role of mailing lists *(academic listserv)* in the Digital Humanities as an example of the close connection between academic communication and Community Building.

43 See https://www.clarin.eu (Accessed: 19 June 2024).
44 See https://www.dariah.eu (Accessed: 19 June 2024). Vgl. Edmond et al. (2020).
45 See https://www.resilience-ri.eu (Accessed: 19 June 2024).
46 See https://roadmap2021.esfri.eu/projects-and-landmarks/browse-the-catalogue/?domain=Social+%26+Cultural+Innovation (Accessed: 19 June 2024).

Mailing lists were adopted very early on in the development of the Digital Humanities given their interdisciplinary nature; the Digital Humanities were, and still are, more *dispersed* than traditional humanities communities. In their early phase (mid-1980s, early 1990s), academic mailing lists in the Digital Humanities were promising for scholarly communication primarily because of their potential to overcome geographical and temporal barriers to bring people together as a "potentially revolutionary replacement for a variety of formal academic communication channels, such as the college classroom, the scholarly journal, the academic conference, and even the scholarly society" (O'Donnell 2020, 185) in a thematically grouped or centrally organized or moderated form. Its second, less common use in the spectrum of Community Building, as an "invisible water cooler" (ibid., 191), is more familiar today. The familiarity is partly due to an adoption by commercial social media platforms like Facebook, Twitter, and Instagram, and partly because it is now standard for academic mailing lists (e.g., originally *Digital Medievalist*[47], *Global Outlook::Digital Humanities*[48]). Here, the mailing list becomes a place for discussion whose members can ask and answer questions, publish announcements and, at least in the early years, hold long and short discussions, debates, and post comments (ibid., 192). This method of communication is comparable to other academic para-discussions (e.g., informal conversations in the hallways of conferences or institutions) (ibid., 194). Even today, mailing lists function as digital "water coolers" and have become the most important tool(s) for academic information dissemination (ibid., 185).

> Academic mailing lists have become a core part of scholarly para- and meta-communication […]. With the advent of the listserv, academics organizing colloquia or conferences, or putting together special collections or journal issues can use the new technology to reach a far wider network of potential participants in a far shorter period of time, including non-members and people outside their immediate circle of acquaintances. While this was rarely identified by the pioneers of the new technology as a potential benefit, it has turned out, in the end, to represent the real revolutionary development, creating a significant improvement in access for marginalized groups and people working outside the main research centres that in many ways represent a far greater disruption of scholarly practice than the early enthusiasts of the listserv-as-journal hoped to create (ibid., 197 f.).

However, the focus is now more on the distribution of information or requests for information since heated discussions often led to complaints (ibid., 196).

In specific ways, online communities are helping to overcome the academic equivalent of the digital divide (ibid., 202). Some online academic communities still

47 See https://journal.digitalmedievalist.org (Accessed: 19 June 2024).
48 See http://www.globaloutlookdh.org (Accessed: 19 June 2024).

using mailing lists, such as the above mentioned, have been complemented over time by academic activities that are not based on email lists or even offline (ibid., 198). In addition, online academic communities emerged on commercial social media platforms to promote social communication (ibid.). Due to increasing criticism of commercial platforms and complicated developments regarding individual platforms (e.g., Twitter or rather X), a new generation of non-commercial online academic communities is currently emerging. In addition to versatile *scholarly networks* with a focus on the humanities, such as *Humanities Commons*[49], or *MLA Commons*[50], subject-specific communities are emerging for the Digital Humanities, such as the Mastodon server *Fedihum*[51].

Despite some overlaps, scholarly societies and online communities have so far had rather different (self-defined) tasks and roles:

1. Scholarly societies: Certification (in the sense of evaluation, assessment, e.g. journals, annual conferences, prizes) and lobbying;[52]
2. Online communities: networking and informal exchange.

For scholarly societies, offering and using non-commercial services for communicating research can make contribute to strengthening the sense of community among members, as well as strengthening their visibility and their specific scientific goals and values, while also strengthening the scholarly led science system in general. According to Fitzpatrick (2020, 351–353), these and other approaches, e.g., the extent to which learned societies can contribute to solving previously unresolved issues in related areas such as research data management (e.g. by creating *social communities* around data), should be pursued further.

Community Building includes both online and offline activities. There is a close connection between Community Building, internal, informal scientific communication and the basic social conditions of the knowledge community:

> Internal, informal science communication structures and organizes the social system of science and its knowledge production. It precedes formal scientific communication, with peers exchanging ideas at workshops, during conference breaks, via telephone or in email discussions. This social network is characterized by a high degree of interaction and the exchange of current information, while at the same time restricting access and making it more difficult to access (Voigt 2012, 17 f.).

49 See https://hcommons.org (Accessed: 19 June 2024).
50 See https://mla.hcommons.org (Accessed: 19 June 2024).
51 See https://fedihum.org/home (Accessed: 19 June 2024).
52 Community Building can also play an important role in scholarly societies.

These claims underline the fact that informal communication in the context of community building is often the basis for the development of new research ideas and collaborations.[53]

4. Training and Education as Task of Community Building

The third aspect of community building is the support of training, education, and mentoring programs to strengthen the digital competencies and skills of scholars in the humanities.[54] By strengthening skills within the community, members can apply digital methods in their own areas of research and practice or become community members. While some specialize in Digital Humanities, others want to acquire the ability to use digital methods without immediately specializing in this field and making it the focus of their research career (Benatti et al. 2021, para. 1). The latter do not yet identify themselves as Digital Humanities scholars (i.e., they are not directly part of the *community of practice*), but still want to and can benefit from digital methods:

> The non-identifying DHer represents a significant proportion of those reaching out for relevant training. Thus it is essential for us to consider how other models – of DH, of pedagogy, and of learning – might support such learners to engage with critical digital humanities practices in a meaningful way (Benatti et al. 2021, para. 1).

The learning needs of different target groups can be met, e.g., by supporting the organization of training courses, workshops, or online courses to promote the interdisciplinary, shared use of digital methods, technologies, and tools, but communities can also be trained using forms and formats, as in the case of the *Programming Historian* initiative.[55] Not only are formats important, but so is the agreement on framework conditions, such as frameworks for digital literacy, scholarship, and mentoring programs or internships (McCarl 2021). These meta-discussions are important building blocks of Community Building and participation in the community of practice in the Digital Humanities, which in turn can be important for the development of a personal learning network in the service of informal continuing education. The existence of

53 Not least for this reason, the loss of personal, informal conversations and encounters during online events was lamented, which led to fewer bonds and the generation of new ideas and collaborations. Systematic subject-specific studies of this phenomenon do not appear to be available.

54 This subsection touches on some aspects of *Digital Humanities pedagogy*. The anthologies Hirsch (2012); Battershill & Ross (2017); and Croxall & Jakacki (2023) should be mentioned here as representative of the extensive literature.

55 See https://programminghistorian.org (Accessed: 19 June 2024).

corresponding offers can therefore be an important incentive to join corresponding informal and formal communities.

5. Cooperation and Projects in the Spirit of Community Building

The fourth and final aspect is the promotion of collaborative projects and initiatives within the community. This aspect of Community Building is discussed separately due to the centrality of collaborative practices in the Digital Humanities – most recently emphasized by the motto "Collaboration as Opportunity" of the global Digital Humanities Conference (DH 2023)[56] – despite overlaps with the previous three aspects.

Collaborative practices and projects include, for example, the joint development of research projects, cooperation in the creation of digital resources, or the organization of joint events. By working together, community members can learn from each other, explore synergies, and achieve common goals. Institutions, research funders, and organizations, such as scholarly societies, can support collaborative initiatives with resources and funding. Grassroots initiatives from the field of Digital Humanities, such as vDHd2021[57] (as a community-driven alternative format for the annual conference of the DHd Association, which was postponed due to the pandemic) or *Saving Ukrainian Cultural Heritage Online* (SUCHO),[58] as a politically motivated global initiative for the preservation of Ukrainian cultural heritage, demonstrate the inherent strength of communities.

While the advantages of collaborative practices in an interdisciplinary field such as the Digital Humanities are apparent, there are also obstacles to overcome. These include, for example, tenure and promotion processes and evaluation criteria (Edmond 2016; Ernst 2015; Huculak & Goddard 2016). Scholarly societies and other community-based interest groups can promote the further development of evaluation criteria of both a general (e.g., DORA[59]) and subject-specific nature (e.g. *MLA Guidelines for Evaluating Work in Digital Humanities and Digital Media*[60]) (see also CoARA[61]).

56 "This year's conference theme 'Collaboration as Opportunity' showcases transdisciplinary and transnational collaboration, with a special focus on the thriving South-Eastern European Digital Humanities community. It will explore how mutual empowerment and collaboration of neighboring countries – regardless of continent and geopolitical placement – can transform regional hubs of expertise to international networks of excellent research, to the benefit of the global DH community" https://dh2023.adho.org (Accessed: 19 June 2024).

57 See https://vdhd2021.hypotheses.org (Accessed: 19 June 2024).

58 https://www.sucho.org (Accessed: 19 June 2024).

59 https://sfdora.org (Accessed: 19 June 2024).

60 https://www.mla.org/About-Us/Governance/Committees/Committee-Listings/Professional-Issues/Committee-on-Information-Technology/Guidelines-for-Evaluating-Work-in-Digital-Humanities-and-Digital-Media (Accessed: 19 June 2024).

61 https://coara.eu (Accessed: 19 June 2024).

6. Conclusion and Outlook

Science communication and Community Building have changed significantly over time due to internationalization, globalization, and digitalization. The socio-technological change described (e.g., the internet, social web, digital research methods, etc.) indicates an increasing reorientation from self-interest to a community of sharing, supported by new technologies and guidelines to promote openness at national and international level, e.g., German Research Foundation (DFG) or European Research Council (ERC). Researchers are increasingly recognizing how they can intrinsically benefit from sharing through greater openness. In turn, not only can academics benefit from better access or greater visibility, embracing open values also has benefits for society and the economy. Digital Humanities scholars are facing challenging opportunities to bring ethical values and technological expertise to discussions and technology impact assessments around current topics such as the use of AI tools through their specific perspective. Doing so reinforces the point that science communication is not only the communication of science but also about communication about science itself.

The paradigm shifts from print to digital and from closed to open are externalizing parts of science communication as well as community building. The desired greater participation of the broader public creates new challenges that should be further investigated, esp. regarding their influence on trust in science (Soderberg et al. 2020).

To achieve innovative, non-commercial, science-appropriate communication of research, new and different formats and infrastructures are needed that do not currently exist or are still in their infancy. What is needed is not only greater acceptance and support for digital, open, and collaborative approaches and more skills development, but also more courage to experiment.

Academics are central to the processes described. However, given their diverse tasks in research, teaching, and professional uncertainties, they often lack the time to engage intensively with the latest developments in science communication and its practice. Although it is advantageous for them to possess a more profound understanding of the scholarly communication system and the associated elements that have been previously discussed, an alternative approach could be to rely on the structures of scholarly infrastructure that are based on the division of labour for specific aspects, in order to navigate the ever-changing landscape (publications, networking, evaluation). Nevertheless, academics should not relinquish responsibility, but should actively participate in the further development of these aspects and be closely involved in the design of corresponding services to promote their usefulness and acceptance and counteract the risk of disruptive innovations.

Quality assurance plays a crucial role in ensuring the credibility and reliability of scholarly results. There is a case for relying less on external proxies to assess quality in the future and instead investing more time and resources in better framework conditions for the open assessment of scientific achievements, e.g., through open access publications, open data and reproducible or more transparent research practices. Sustainable and FAIR access to scientific data and research results is an important prerequisite for preserving valuable information for the future in a reusable form. Scholars should strive to regain control over their own workflows and means of communication or develop alternative approaches such as *scholarly-led* publishing. In this respect, scientific policy and academic institutions have a major responsibility for support and funding these efforts.

Bibliography

acatech (2017). *Social Media und digitale Wissenschaftskommunikation. Analyse und Empfehlungen zum Umgang mit Chancen und Risiken in der Demokratie.* München/Halle/Mainz: acatech – Deutsche Akademie der Technikwissenschaften | Deutsche Akademie der Naturforscher Leopoldina | Union der deutschen Akademien der Wissenschaften. URL: https://www.leopoldina.org/uploads/tx_leopublication/2017_Stellungnahme_WOeM_web.pdf (Accessed: 19 June 2024).

AG Digitales Publizieren (2021). Digitales Publizieren in den Geisteswissenschaften. Begriffe, Standards, Empfehlungen, *Zeitschrift für digitale Geisteswissenschaften. Working Papers,* 1, no pag. DOI: https://doi.org/10.17175/wp_2021_001 (Accessed: 19 June 2024).

Balula, A., & Leão, D. (2021). Multilingualism within Scholarly Communication in SSH. A literature review, *Italian Journal of Library, Archives and Information Science,* 12(2), 88–98. DOI: https://doi.org/10.4403/jlis.it-12672 (Accessed: 19 June 2024).

Battershill, C., & Ross, Sh. (2017). *Using digital humanities in the classroom. A practical introduction for teachers, lecturers and students.* London/Oxford/New York: Bloomsbury.

Baumgärtner, I., Kern, M., Leven, K.-H., & Toepfer, R. (2021). Mittelalter erschließen, *Das Mittelalter. Perspektiven mediävistischer Forschung* 26(1), 68–86. DOI: https://doi.org/10.17885/heiup.mial.2021.1.24310 (Accessed: 19 June 2024).

Becerril-García, A. (2019). AmeliCA vs Plan S. Same target, two different strategies to achieve Open Access. *Voces AmeliCA* [Blog]. URL: http://amelica.org/index.php/en/2019/02/10/amelica-vs-plan-s-same-target-two-different-strategies-to-achieve-open-access (Accessed: 19 June 2024).

Benatti, F., Gooding, P., & Sillence, M. (2021). Learning Digital Humanities in a Community of Practice. The DEAR model of Postgraduate Research Training, *Digital Humanities Quarterly,* 15(3), 1–77. URL: http://www.digitalhumanities.org/dhq/vol/15/3/000570/000570.html (Accessed: 19 June 2024).

Borek, L., Hastik, C., Khramova, V., Illmayer, K., & Geiger, J.D. (2021). Information Organization and Access in Digital Humanities. TaDiRAH Revised, Formalized and FAIR. In T. Schmidt & C. Wolff (Eds.), *Information between Data and Knowledge* (pp. 321–332). Glückstadt: Werner Hülsbusch [= *Schriften zur Informationswissenschaft,* 74]. DOI: https://doi.org/10.5283/epub.44951 (Accessed: 19 June 2024).

Brembs, B., Lenardic, A., Murray-Rust, P., Chan, L., & Irawan, D.E. (2023). Mastodon over Mammon. towards publicly owned scholarly knowledge, *Royal Society Open Science,* 10(230207), 1–9. DOI: https://doi.org/10.1098/rsos.230207 (Accessed: 19 June 2024).

Brembs, B., & Siems, R. (2023). Die Wiederkehr der Wunderkammer. Bibliothek 2040 – eine Utopie, *Bibliothek Forschung und Praxis,* 47(1), 22–25. DOI: https://doi.org/10.1515/bfp-2022-0061 (Accessed: 19 June 2024).

Burghardt, Manuel (2020). Theorie und Digital Humanities. Eine Bestandsaufnahme. *AG Digital Humanities Theorie* [Blog]. URL: https://dhtheorien.hypotheses.org/680 (Accessed: 19 June 2024).

Id., Czmiel, A., Dieckmann, L., Guhr, S., Jacke, J., Reiter, N., Scholger, W., & Wuttke, U. (2022). Offen für alle(s)? Open Identities im Reviewprozess der DHd-Konferenz. In M. Geierhos, P. Trilcke, I. Börner, S. Seifert, A. Busch, & P. Helling (Eds.), *DHd 2022. Kulturen des digitalen Gedächtnisses. Konferenzabstracts* (pp. 21–24). Potsdam: Verband “Digital Humanities im deutschsprachigen Raum”. DOI: https://doi.org/10.5281/zenodo.6304590 (Accessed: 19 June 2024).

Burghardt, M., & Wolff, Ch. (2015). Zentren für Digital Humanities in Deutschland, *Information – Wissenschaft & Praxis,* 66(5–6), 312–326. DOI: https://doi.org/10.1515/iwp-2015-0056 (Accessed: 19 June 2024).

Busch, A., Meister, J.Ch., & Schumacher, M. (2016). Wo bleibt eigentlich der einzelne Fachwissenschaftler? Community Building als Aufgabe und Herausforderung für DH-Infrastrukturen, *Bibliothek Forschung und Praxis,* 40(2), 278–282. DOI: https://doi.org/10.1515/bfp-2016-0028 (Accessed: 19 June 2024).

Chen, Ch. (2003). *Mapping Scientific Frontiers. The Quest for Knowledge Visualization.* London: Springer London. DOI: https://doi.org/10.1007/978-1-4471-5128-9 (Accessed: 19 June 2024).

Couldry, N., & Mejias, U.A. (2019). *The costs of connection. How data is colonizing human life and appropriating it for capitalism.* Stanford, California: Stanford University Press [= *Culture and Economic Life*].

Croxall, B., & Jakacki, D.K. (Eds.). (2023). *What we teach when we teach DH. Digital humanities in the classroom.* Minneapolis, MN: University of Minnesota Press [= *Debates in the digital humanities*].

Dang, S.-M. (2017). Zenodo, SocArXiv oder GitHub? Preprints in den Geisteswissenschaften. *oa books* [Blog]. URL: https://www.oabooks.de/zenodo-socarxiv-oder-github-preprints-in-den-geisteswissenschaften (Accessed: 19 June 2024).

Davis, Ph. (2017). Scientific Reports Overtakes PLOS ONE As Largest Megajournal. *The Scholarly Kitchen* [Blog]. URL: https://scholarlykitchen.sspnet.org/2017/04/06/scientific-reports-overtakes-plos-one-as-largest-megajournal (Accessed: 19 June 2024).

De Silva, P.U.K., & Vance, C. (2017). *Scientific scholarly communication. The changing landscape.* New York/Berlin/Heidelberg: Springer [= *Fascinating Life Sciences*].

Dernbach, B., Kleinert, Ch., & Münder, H. (2012). Einleitung. Die drei Ebenen der Wissenschaftskommunikation (pp. 1–15). In eid. (Eds.), *Handbuch Wissenschaftskommunikation.* Wiesbaden: Springer VS. DOI: https://doi.org/10.1007/978-3-531-18927-7 (Accessed: 19 June 2024).

Dulong De Rosnay, M., & Stalder, F. (2020). Digital commons, *Internet Policy Review,* 9(4), no pag. DOI: https://doi.org/10.14763/2020.4.1530 (Accessed: 19 June 2024).

Edmond, J. (2016). Collaboration and Infrastructure. In S. Schreibman, R. Siemens & J. Unsworth (Eds.), *A new Companion to Digital Humanities* (pp. 54–65). Chichester, West Sussex: Wiley-Blackwell [= *Blackwell Companions to Literature and Culture,* 93].

Ead. (Ed.). (2020). *Digital Technology and the Practices of Humanities Research.* Cambridge, United Kingdom: Open Book Publishers. DOI: https://doi.org/10.11647/OBP.0192 (Accessed: 19 June 2024).

Ead., Fischer, F., Tasovac, T., & Romary, L. (2020). Springing the Floor for a Different Kind of Dance. Building DARIAH as a Twenty-First Century Research Infrastructure for the Arts and Humanities. In J. Edmond (Ed.), *Digital Technology and the Practices of Humanities Research* (pp. 207–234). Cambridge, United Kingdom: Open Book Publishers. DOI: https://doi.org/10.11647/OBP.0192 (Accessed: 19 June 2024).

Edmond, J., & Romary, L. (2020). Academic Publishing. New Opportunities for the Culture of Supply and the Nature of Demand. In J. Edmond (Ed.), *Digital Technology and the Practices of Humanities Research* (pp. 49–80). Cambridge, United Kingdom: Open Book Publishers. DOI: https://doi.org/10.11647/OBP.0192 (Accessed: 19 June 2024).

Ernst, Th. (2015). Vom Urheber zur Crowd, vom Werk zur Version, vom Schutz zur Öffnung? Kollaboratives Schreiben und Bewerten in den Digital Humanities, *Zeitschrift für digitale Geisteswissenschaften,* Sonderband 1, no pag. DOI: https://doi.org/10.17175/sb001_021 (Accessed: 19 June 2024).

Estill, L., Guiliano, J., Ortega, É., Terras, M., Verhoeven, D., & Layne-Worthey, G. (2022). The circus we deserve? A front row look at the organization of the annual academic conference for the Digital Humanities, *Digital Humanities Quarterly,* 16(4), 1–79. URL: http://www.digitalhumanities.org/dhq/vol/16/4/000643/000643.html (Accessed: 19 June 2024).

Eve, M.P., & Gray, J. (Eds.). (2020). *Reassembling scholarly communications. histories, infrastructures, and global politics of open access.* Cambridge, Massachusetts:

The MIT Press. DOI: https://doi.org/10.7551/mitpress/11885.001.0001 (Accessed: 19 June 2024).
Fecher, B. (2022). Welche Denkfehler die gesellschaftliche Relevanz von Forschung in Deutschland hemmen, *Transfer & Innovation,* 1, 123–137. URL: https://www.transfer-und-innovation.de/media/gnnboh/ygowc5/ca5826736c0e0002cfb1bc083057a1966f25ca69.pdf (Accessed: 19 June 2024).
Fenlon, K., Reza, A., Grimmer, J., & Wagner, T. (2023). Community-centric factors in sustaining digital scholarship. In A. Baillot, W. Scholger, T. Tasovac, & G. Vogeler (Eds.), *Digital Humanities 2023. Book of Abstracts* (pp. 273–275). Graz: Zentrum für Informationsmodellierung. Austrian Center for Digital Humanities | Universität Graz. DOI: https://doi.org/10.5281/zenodo.7961822 (Accessed: 19 June 2024).
Fitzpatrick, K. (2020). Not all Networks. Toward Open, Sustainable Research Communities. In M. P. Eve & J. Gray (Eds.), *Reassembling scholarly communications. histories, infrastructures, and global politics of open access* (pp. 351–359). Cambridge, Massachusetts: The MIT Press. DOI: https://doi.org/10.7551/mitpress/11885.003.0035 (Accessed: 19 June 2024).
Frick, C., & Seltmann, M. (2023). Referenzrahmen für eigenständige digitale Wissenschaftskommunikation durch Forschende, *Zeitschrift für digitale Geisteswissenschaften. Working Papers,* 3, no pag. DOI: https://doi.org/10.17175/wp_2023b (Accessed: 19 June 2024).
Eaed. (2024). Wissenschaftskommunikation. In U. Engelkenmeier, K. Keller-Loibl, B. Schmid-Ruhe, & R. Stang (Eds.), *Handbuch Bibliothekspädagogik* (pp. 291–304). Berlin/Boston: De Gruyter Saur. DOI: https://doi.org/10.1515/9783111032030-027 (Accessed: 19 June 2024).
Geier, A., & Gottschling, M. (2019). Wissenschaftskommunikation auf Twitter? Eine Chance für die Geisteswissenschaften!, *Mitteilungen des Deutschen Germanistenverbandes,* 66(3), 282–291. DOI: https://doi.org/10.14220/mdge.2019.66.3.282 (Accessed: 19 June 2024).
Gundermann, Ch., Brauer, J., Carlà-Uhink, F., Keilbach, J., Logge, Th., Morat, D., Peselmann, A., Samida, S., Schwabe, A., Sénécheau, M., & G. Koch et al. (2021). *Schlüsselbegriffe der Public History.* Göttingen: Vandenhoeck & Ruprecht [= *utb Geschichte,* 5728].
Hagenhoff, S., Seidenfaden, L., Ortelbach, B., & Schumann, M. (2007). *Neue Formen der Wissenschaftskommunikation. Eine Fallstudienuntersuchung.* Göttingen: Universitätsverlag Göttingen [= *Göttinger Schriften zur Internetforschung,* 4]. DOI: https://doi.org/10.17875/gup2007-208 (Accessed: 19 June 2024).
Hecker, S., Muki, H., Bowser, A., Makuch, Z., Vogel, J., & Bonn, A. (Eds.). (2018). *Citizen Science. Innovation in Open Science, Society and Policy.* London: UCL Press. URL: http://library.oapen.org/handle/20.500.12657/28178 (Accessed: 19 June 2024).

Heise, Ch. (2018). *Von Open Access zu Open Science. Zum Wandel digitaler Kulturen der wissenschaftlichen Kommunikation.* Lüneburg: meson press. URL: http://meson.press/books/von-open-access-zu-open-science (Accessed: 19 June 2024).

Hirsch, B.D. (Ed.). (2012). *Digital Humanities Pedagogy. Practices, Principles and Politics.* Cambridge: Open Book Publishers. DOI: https://doi.org/10.11647/OBP.0024 (Accessed: 19 June 2024).

Howard-Sukhil, Ch., Wallace, S., & Chakrabarti, A. (2021). Developing Research through Podcasts. Circulating Spaces, A Case Study, *Digital Humanities Quarterly,* 15(3), no pag. URL: http://www.digitalhumanities.org/dhq/vol/15/3/000554/000554.html (Accessed: 19 June 2024).

Huculak, J.M., & Goddard, L. (2016). Is Promotion and Tenure Inhibiting DH/Library Collaboration? A Case for Care and Repair. *dh + lib* [Blog]. URL: http://acrl.ala.org/dh/2016/07/29/a-case-for-care-and-repair (Accessed: 19 June 2024).

Klein, J.Th. (2015). *Interdisciplining Digital Humanities. Boundary Work in an Emerging Field.* Digital Humanities. Ann Arbor: University of Michigan Press. DOI: https://doi.org/10.3998/dh.12869322.0001.001 (Accessed: 19 June 2024).

Kleineberg, M., & Kaden, B. (2017). Open Humanities? ExpertInnenmeinungen über Open Access in den Geisteswissenschaften, *LIBREAS. Library Ideas,* 32, no pag. URL: https://libreas.eu/ausgabe32/kleineberg (Accessed: 19 June 2024).

König, M. (2015). Herausforderung für unsere Wissenschaftskultur. Weblogs in den Geisteswissenschaften. In W. Schmale (Ed.), *Digital Humanities. Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität* (pp. 57–74). Stuttgart: Franz Steiner Verlag [= *Historische Mitteilungen – Beihefte,* 91]. DOI: https://doi.org/10.25162/9783515111508 (Accessed: 19 June 2024).

Ead. (2022) Tröten über Droysen. Ein Mastodon-Leitfaden für Historiker:innen. *Digital Humanities am DHIP* [Blog]. URL: https://dhdhi.hypotheses.org/7205 (Accessed: 19 June 2024).

Könneker, C. (2020). Wissenschaftskommunikation und Social Media. Neue Akteure, Polarisierung und Vertrauen. In J. Schnurr & A. Mäder (Eds.), *Wissenschaft und Gesellschaft. Ein vertrauensvoller Dialog* (pp. 25–47). Berlin/Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-662-59466-7_3 (Accessed: 19 June 2024).

Kunz, R. (2022). Threats to Academic Freedom under the Guise of Open Access. The Power of Publishers, Data Tracking in Science, and the Responsibilities of Public Actors. *Verfassungsblog* [Blog]. URL: https://verfassungsblog.de/threats-to-academic-freedom-under-the-guise-of-open-access (Accessed: 19 June 2024).

Larivière, V., Haustein, S., & Mongeon, Ph. (2015). The Oligopoly of Academic Publishers in the Digital Era, *PLOS ONE,* 10(6), 1–15. DOI: https://doi.org/10.1371/journal.pone.0127502 (Accessed: 19 June 2024).

Maggio, L.A., Jeffrey, A., Haustein, S., & Samuel, A. (2022). Becoming metrics literate. An analysis of brief videos that teach about the h-index, *PLOS ONE,* 17(5), 1–16. DOI: https://doi.org/10.1371/journal.pone.0268110 (Accessed: 19 June 2024).

McCarl, C. (2021). An Approach to Designing Project-Based Digital Humanities Internships, *Digital Humanities Quarterly,* 15(3), 1–76. URL: http://www.digitalhumanities.org/dhq/vol/15/3/000567/000567.html (Accessed: 19 June 2024).

Neuber, F., & Sahle, P. (2022). Nach den Büchern. Rezensionen digitaler Forschungsressourcen, *H-Soz-Kult. Forum. Buchrezensionen in den Geschichtswissenschaften,* no pag. URL: www.hsozkult.de/debate/id/fddebate-132457 (Accessed: 19 June 2024).

Neuroth, H. (2017). Bibliothek, Archiv, Museum. In F. Jannidis, H. Kohle & M. Rehbein (Eds.), *Digital Humanities. Eine Einführung* (pp. 213–222). Stuttgart: J.B. Metzler Verlag.

Niemann, Ph., Schrögel, Ph., & Hauser, Ch. (2017). Präsentationsformen der externen Wissenschaftskommunikation. Ein Vorschlag zur Typologisierung, *Zeitschrift für Angewandte Linguistik,* 66, 81–113. DOI: https://doi.org/10.1515/zfal-2017-0019 (Accessed: 19 June 2024).

Niemann, Ph., Van Den Bogaert, V., & Ziegler, R. (Eds.). (2023). *Evaluationsmethoden der Wissenschaftskommunikation.* Wiesbaden: Springer Fachmedien. DOI: https://doi.org/10.1007/978-3-658-39582-7 (Accessed: 19 June 2024).

Nyhan, J. (2020). The Evaluation and Peer Review of Digital Scholarship in the Humanities. Experiences, Discussions, and Histories. In J. Edmond (Ed.), *Digital Technology and the Practices of Humanities Research* (pp. 163–181). Cambridge, United Kingdom: Open Book Publishers. DOI: https://doi.org/10.11647/OBP.0192 (Accessed: 19 June 2024).

O'Donnell, P.P. (2020). Critical Mass. The Listserv and the Early Online Community as a Case Study in the Unanticipated Consequences of Innovation in Scholarly Communication. In J. Edmond (Ed.), *Digital Technology and the Practices of Humanities Research* (pp. 183–206). Cambridge, United Kingdom: Open Book Publishers. DOI: https://doi.org/10.11647/OBP.0192 (Accessed: 19 June 2024).

Pasternack, P. (2022). *Wissenschaftskommunikation, neu sortiert. Eine Systematisierung der externen Kommunikationen der Wissenschaft.* Wiesbaden: Springer Fachmedien.

Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship, *Quantitative Science Studies,* 1(1), 428–444. DOI: https://doi.org/10.1162/qss_a_00023 (Accessed: 19 June 2024).

Peters, I. (2023). Social Media & Social Web. In R. Kuhlen, D. Lewandowski, W. Semar, & Ch. Womser-Hacker (Eds.), *Grundlagen der Informationswissenschaft* (pp. 691–702). Berlin/Boston: De Gruyter Saur. DOI: https://doi.org/10.1515/9783110769043-059 (Accessed: 19 June 2024).

Prescott, A. (2016). Beyond the Digital Humanities Center. The administrative landscapes of the Digital Humanities. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A new Companion to Digital Humanities* (pp. 461–475). Chichester, West Sussex: Wiley-Blackwell [= *Blackwell Companions to Literature and Culture,* 93].

Roeder, T., Söring, S., Dogunke, S., Elwert, F., Wübbena, Th., Lordick, H., Cremer, F., & Klammt, A. (2019). Digital Humanities “from Scratch”. Ein Panel-Bericht zur DHd 2019. #DhfromScratch. #dhd2019. *DHd Blog* [Blog]. URL: https://dhd-blog.org/?p=11804 (Accessed: 19 June 2024).

Rosendaal, H.E., & P.A.Th.M. Geurts (1999). Forces and functions in scientific communication. An analysis of their interplay. *Paper presented at Conference on Co-operative Research in Information Systems in Physics* (pp. 1–32). Enschede: University of Twente. URL: https://ris.utwente.nl/ws/portalfiles/portal/6162491/Roosendaal97forces.pdf (Accessed: 19 June 2024).

Ross-Hellauer, T. (2017). What is open peer review? A systematic review. *F1000Research,* 6(588), 1–39. DOI: https://doi.org/10.12688/f1000research.11369.2 (Accessed: 19 June 2024).

Sahle, P. (2015). Digital Humanities? Gibt’s doch gar nicht! In C. Baum & Th. Stäcker (Eds.), *Grenzen und Möglichkeiten der Digital Humanities.* Wolfenbüttel: Herzog August Bibliothek [= *Zeitschrift für digitale Geisteswissenschaften.* Sonderband, 1]. DOI: https://doi.org/10.17175/sb001_004 (Accessed: 19 June 2024).

Schöch, Ch. (2016). Warum Academia.edu? Eine Replik. *Romanische Studien,* 4 [Blog]. URL: http://blog.romanischestudien.de/warum-academia-edu-eine-replik (Accessed: 19 June 2024).

Schröder, S. (2023). “Wir haben Angst, die falschen Fragen zu stellen”. *Wissenschaftskommunikation.de* [Blog]. URL: https://www.wissenschaftskommunikation.de/chatgpt-in-der-wissenschaftskommunikation-66823 (Accessed: 19 June 2024).

Schuldt-Baumgart, N. (2022). Was Wissenstransfer und Wissenschaftskommunikation unterscheidet. *Wissenschaftskommunikation.de* [Blog]. URL: https://www.wissenschaftskommunikation.de/was-wissenstransfer-und-wissenschaftskommunikation-unterscheidet-58417 (Accessed: 19 June 2024).

Schwan, A., & Thomson, T. (Eds.). (2022). *The Palgrave handbook of digital and public humanities.* Cham: Palgrave Macmillan.

Seltmann, M.E.-H. (2023). #PublicDH oder doch nur #WissKomm? In P. Trilcke, A. Busch & P. Helling (Eds.), *DHd 2023. Open Humanities Open Culture* (pp. 1–5). Trier/Luxemburg: Zenodo. DOI: https://doi.org/10.5281/zenodo.7715494 (Accessed: 19 June 2024).

Shi, J. (2023). Articulations of language and value(s) in scholarly publishing circuits, *Canadian Journal of Academic Librarianship,* 9, 1–33. DOI: https://doi.org/10.33137/cjal-rcbu.v9.38148 (Accessed: 19 June 2024).

Soderberg, C.K., Errington, T., M., & Nosek, B.A. (2020). Credibility of preprints. An interdisciplinary survey of researchers, *Royal Society Open Science,* 7(10.201520), no pag. DOI: https://doi.org/10.1098/rsos.201520 (Accessed: 19 June 2024).

Söllner, K. (2017). Geisteswissenschaften. In B. Mittermaier & ead. (Eds.), *Praxishandbuch Open Access* (pp. 247–253). Berlin/Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110494068 (Accessed: 19 June 2024).

Toledo, E. G. (2020). Why Books are Important in the Scholarly Communication System in Social Sciences and Humanities, *Scholarly Assessment Reports,* 2(1.6), 1–8. DOI: https://doi.org/10.29024/sar.14 (Accessed: 19 June 2024).

Umlauf, K., & Gradmann, S. (2014). Publikationskreislauf. In eid. (Eds.), *Lexikon der Bibliotheks- und Informationswissenschaft,* 2. K–Z. (pp. 968–969). Stuttgart: Anton Hiersemann.

Van der Weel, A., & Praal, F. (2020). Publishing in the Digital Humanities. The Treacle of the Academic Tradition. In J. Edmond (Ed.), *Digital Technology and the Practices of Humanities Research* (pp. 21–48). Cambridge, United Kingdom: Open Book Publishers. DOI: https://doi.org/10.11647/OBP.0192 (Accessed: 19 June 2024).

Voigt, K. (2012). *Informelle Wissenschaftskommunikation und "Social Media".* Berlin: Frank & Timme [= *Kulturen – Kommunikation – Kontakte,* 10].

Weitze, M.-D., & Heckl, W. M. (2016). *Wissenschaftskommunikation. Schlüsselideen, Akteure, Fallbeispiele.* Berlin/Heidelberg: Springer Spektrum. DOI: https://doi.org/10.1007/978-3-662-47843-1 (Accessed: 19 June 2024).

Winters, J. (2020). Learned Societies, Humanities Publishing, and Scholarly Communication in the UK. In M. P. Eve & J. Gray (Eds.), *Reassembling scholarly communications. histories, infrastructures, and global politics of open access* (pp. 343–349). Cambridge, Massachusetts: The MIT Press. DOI: https://doi.org/10.7551/mitpress/11885.003.0034 (Accessed: 19 June 2024).

Wróblewska, M. N. (2021). Research impact evaluation and academic discourse. *Humanities and Social Sciences Communications,* 8(1.58), 1–12. DOI: https://doi.org/10.1057/s41599-021-00727-8 (Accessed: 19 June 2024).

Wuttke, U. (2022). Wege bereiten, vermitteln und Denkräume schaffen! Reflexionen zu institutionellen und infrastrukturellen Erfolgsfaktoren für Digital Humanities an deutschen Universitäten auf Grundlage von Expert*innen-interviews, *Zeitschrift für digitale Geisteswissenschaften,* 7, 1–58. DOI: https://doi.org/10.17175/2022_006 (Accessed: 19 June 2024).

Ead., & Gebert, B. (2021). How to make your medieval research more visible with open scholarship methods and tools, *Imago Temporis. Medium Aevum,* 15, 415–450. DOI: https://doi.org/10.21001/itma.2021.15.14 (Accessed: 19 June 2024).

Ziegler, R., & Fischer, L. (2020). *Ziele von Wissenschaftskommunikation. Eine Analyse der strategischen Ziele relevanter Akteure für die institutionelle Wissenschaftskommunikation in Deutschland, 2014–2020.* Berlin: Wissenschaft im Dialog. URL: https://impactunit.de/wp-content/uploads/2021/08/Ergebnisbericht_Ziele_von_Wisskomm.pdf (Accessed: 19 June 2024).

Further References

AWBI (2021). *Datentracking in der Wissenschaft. Aggregation und Verwendung bzw. Verkauf von Nutzungsdaten durch Wissenschaftsverlage. Ein Informationspapier des Ausschusses für Wissenschaftliche Bibliotheken und Informationssysteme der Deutschen Forschungsgemeinschaft.* Zenodo [Information Paper]. DOI: https://doi.org/10.5281/zenodo.5900759 (Accessed: 19 June 2024).

Balula, A., Caliman, L., Costa, C., Fiorini, S., Jarmelo, S., Leão, D., Mornier, P., Nomine, J.-F., Padez, M. J., Plag, C., Pölönen, J., del Rio Rande, G., Silva, M., Stojanovski, J., Szulińska, A., Volt, I., Giménez Toledo, E., & Withanage, D. (2021). OPERAS Multilingualism White Paper (SIG Multilingualism). OPERAS [Website]. URL: https://operas-eu.org/special-interest-group-living-book/operas-multilingualism-white-paper-june-2021 (Accessed: 19 June 2024).

BMBF (2019). Grundsatzpapier des Bundesministeriums für Bildung und Forschung zur Wissenschaftskommunikation. Bundesministerium für Bildung und Forschung. URL: https://www.bmbf.de/bmbf/de/ueber-uns/wissenschaftskommunikation-und-buergerbeteiligung/wissenschaftskommunikation/wissenschaftskommunikation.html (Accessed: 19 June 2024).

Bosman, J., & Kramer, B. (2018). Workflows, *Innovations in Scholarly Communication. Changing Research Workflows.* Universität Utrecht [Website]. URL: https://101innovations.wordpress.com/workflows (Accessed: 19 June 2024).

DARIAH-EU (2018). Towards a Plan(HS)S. DARIAH's position on PlanS. Berlin [Position Paper]. URL: https://www.dariah.eu/wp-content/uploads/2018/10/Towards-a-PlanHSS-excerpt.pdf (Accessed: 19 June 2024).

Frick, C., Heller, L., Ramünke, S., & Strauß, F. (2021). Bibliotheken als Dienstleisterinnen und Labore der Wissenschaftskommunikation. #vBIB21. Zenodo [Workshop-Slides]. DOI: https://doi.org/10.5281/zenodo.5752401 (Accessed: 19 June 2024).

Unsworth, J. (2000). Scholarly Primitives. What methods do humanities researchers have in common, and how might our tools reflect this? *[Website of the Author].* URL: http://johnunsworth.name/Kings.5-00/primitives.html (Accessed: 19 June 2024).

Id. (2003). Not-so-modest Proposals. What do we want our system of scholarly communication to look like in 2010? (CIC Summit on Scholarly Communication). *[Website of the Author].* URL: http://johnunsworth.name/CICsummit.htm (Accessed: 19 June 2024).

Wikipedia (2023). Fediverse [status: 18 November 2023]. *Wikipedia – Die freie Enzyklopädie.* URL: https://de.wikipedia.org/w/index.php?title=Fediverse&oldid=239231731 (Accessed: 19 June 2024).

Wissenschaftsrat (2016). Wissens- und Technologietransfer als Gegenstand institutioneller Strategien [Position Paper]. URL: https://www.wissenschaftsrat.de/download/archiv/5665-16.pdf?__blob=publicationFile&v=1 (Accessed: 19 June 2024).
Id. (2021). Wissenschaftskommunikation [Position Paper]. URL: https://www.wissenschaftsrat.de/download/2021/9367-21.html (Accessed: 19 June 2024).

Figure Credits

Both figures were reconstructed and translated by the editors.

Fig. 1: Traditional systematization of science communication, source: acatech 2017, 21.

Fig. 2: Fourfold matrix of science communication, source: Seltmann 2023, 2.

Virtual Research Environments

Caroline T. Schroeder

https://orcid.org/0000-0001-9543-0692

Abstract Virtual Research Environments in theological studies (and esp. early Christian studies and the related field of Classical studies) can provide valuable infrastructure for producing digital editions of primary sources and enabling other forms of digital and computational research. Creating and sustaining these environments has challenges. This chapter examines the benefits of collaborating across projects as well as sharing and reusing digital resources. The chapter also presents some of the considerations for working with *messy* or *clean* digital data, and for adopting existing technical standards. With respect to all of these issues, building and using VREs involves developing relevant technical infrastructure. But just as important as technology are the humanistic questions and collaborative personal relationships underpinning a successful digital initiative.

Keywords Digital Humanities, Virtual Research Environments, Tools, Standards, Collaboration, Open Access, Data Cleaning, Early Christian Studies

1. Introduction

Virtual Research Environments (VRE) in theological studies (and esp. early Christian studies and the related field of Classical studies) can provide valuable infrastructure for producing digital editions of primary sources and enabling other forms of digital and computational research. Creating and sustaining these environments has challenges. Some key elements of successful VREs include collaboration across different projects, sharing and reusing digital resources, and careful consideration of how to work with *messy* or *clean* digital data, and whether to adopt existing technical standards. In this paper, I will address these aspects of Digital Humanities work in our field through the lens of the creation of the *Coptic Scriptorium* (CS) platform. While the focus of this essay is CS, other VREs are considered and the analysis extends beyond the scope of our individual experience.

The virtual research environment I co-direct, *Coptic Scriptorium*, originated at a *National Endowment for the Humanities* summer research institute hosted by the *Perseus Digital Library* at Tufts University in 2012. Researchers from all career stages – from graduate student to full professor – working in a variety of languages – Greek, Latin, Russian, Coptic – applied and attended a three-week workshop co-directed by

Caroline T. Schroeder: Virtual Research Environments, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 379–392. DOI: https://doi.org/10.11588/heibooks.1521.c21960

Monica Berti (a Classicist and Digital Humanist, now of Leipzig University), Gregory Crane (Tufts University, *Perseus* founder), and Anke Lüdeling (Corpus Lingustics, Humboldt University). At that point, "digital Coptic" was in its infancy, and there were few openly accessible VREs for early Christian studies or ancient studies. The *Perseus Digital Library*[1], our institutional host, was one of the most well-known (Crane 1998). *Trismegistos*[2] served as a *linked-data* hub for people, places, and ancient texts (building on and collaborating with the *Heidelberger Gesamtverzeichnis griechischer Urkunden aus Ägypten* [HGV] and the *Leuven Database of Ancient Books* [LDAB]) (Depauw & Gheldof 2014). *Papyri.info* created a cutting-edge collaborative text-editing environment that benefitted from crowd-sourcing among papyrologists.[3] The *Tesserae Project* at the University of Buffalo had also launched, facilitating research in text-reuse in classical sources (Forstall et al. 2011; Okuda et al. 2022; cf. also the chapter by J. Nantke in this volume, p. 288). While there were additional subscription-based research environments for Greek and Latin, open access or open-source environments were few – those aforementioned being some of the main projects. The organizers of the NEH Institute hoped its participants would be inspired to fill the gaps.

In Coptic Studies, the Unicode character set for the Coptic alphabet[4] had been approved in 2004, with major additions in subsequent years, including important diacritics such as the binding macron characters and the "binding ni" that appears at the end of lines in manuscripts in 2007.[5] *Papyri.info* had recently begun publishing some Coptic papyri and ostraca. Other institutes and individuals had been working on both Coptic and Syriac texts non-Unicode fonts and circulating digital forms of the New Testament and Christian Old Testament in these languages (Schroeder 2019). Additionally, Tito Orlandi's work at the *Corpus dei Manuscritti Copti Letterari* (CMCL), ongoing for decades, was foundational (Orlandi 1997a; b; 2021). Nonetheless, sustainable digital editions of Coptic literature and sustained digital and computational research in Coptic studies were only at the beginning stages. Amir Zeldes, a linguist at Humboldt University (not a "Coptologist") and I (not a linguist), met at the Tufts NEH Institute and discovered our shared interest in Coptic literature and Digital Humanities, and began planning the project. *Coptic Scriptorium*[6] launched its first

1 See http://www.perseus.tufts.edu (Accessed: 25 June 2024).

2 See http://www.trismegistos.org (Accessed: 25 June 2024).

3 See http://papyri.info/ddbdp (Accessed: 25 June 2024).

4 See https://www.unicode.org/wg2/docs/n2824.pdf (Accessed: 25 June 2024).

5 For the revisions in 2004, see the worksheet under https://www.unicode.org/wg2/docs/n2744.pdf; for the revisions in 2007 see http://unicode.org/wg2/docs/n3222 and https://www.unicode.org/L2/L2007/07118.htm [Protocol of the UTC 111/L 2 208 Joint Meeting]. For the standard Unicode font *Antinoou* (2012) see http://www.evertype.com/fonts/coptic. All addresses were accessed on 25 June 2024.

6 See https://copticscriptorium.org (Accessed: 25 June 2024).

pilot corpus, natural language processing tools, and one-page website in 2013.[7] We now have a database of Coptic literature of over 1.2 million words (annotated for part of speech, syntax, entities, lemmas, language of origin, manuscript information, and more), plus multiple tools including an online natural language processing pipeline (Schroeder & Zeldes 2013–2023; 2016; 2020).

In this paper, I will address three key issues in developing VREs that have posed both challenges and opportunities as our project has grown over the past decade: specialization and collaboration in reuse of data and tools, messy data, and technical standards. While building and using VREs for Digital Humanities research involves developing technical infrastructure, just as important are pursuing the humanistic questions and collaborative personal relationships underpinning a successful digital initiative.

2. Specialization, Collaboration, and Reuse

Digital research environments are expensive endeavors, and often the audience or user-community for such environments is small. In Coptic Studies, for example, most of us know each other, whether we work in North America, Europe, Australia, Egypt, or Japan. And there is little room for overlap in research – if we already know someone is working on an edition of certain manuscripts or papyri, the rest of us usually go off to work on something else. This also has proven particularly true in digital Coptic Studies. The ecosystem that has emerged consists of specialists in specific areas. And while early Christian studies, Classics, and Biblical Studies have more expansive communities, in the digital realm, the costs of creating VREs discourages duplication. Thus, in textual and linguistic Coptic studies, the major open access projects specialize in different aspects of the field. Each of these research environments has developed in response to the particular research needs of a certain research community, and each has limitations as well as benefits.

Papyri.info publishes digital editions of ostraca and papyri using the XML *(extensible markup language)* standards created by the *Text Encoding Initiative* and the EpiDoc subcommunity of the TEI (Elliott et al. 2006–2021).[8] Papyri and ostraca tend to be shorter than literary texts, and *Papyri.info* creates a digital research environment comparable to the analog research methods papyrologists traditionally have employed (editions and translations with notes, images, apparatus, etc.). As a result,

7 Although we no longer have a copy of the original website, the version from 9 October 2014 is archived at the Internet Archive Wayback Machine: https://web.archive.org/web/20141009102742/http://www.copticscriptorium.org (Accessed: 25 June 2024).

8 See on TEI: http://www.tei-c.org; see on EpiDoc: http://epidoc.stoa.org. Both addresses were accessed on 25 June 2024.

crowd-sourcing digitization of papyri among papyrologists has been feasible. Certainly, *Papyri.info* has invested a significant amount of time and resources in training and outreach, which cannot be understated; the genre of sources and the digital methods also contribute to its success in publishing a tremendous number of documents. There are some features, however, that this environment either does not have or has some challenges with (None of the following observations should be taken as criticisms – it is a remarkable achievement in scope and method. The description of the platform's parameters exemplifies how this particular VRE serves specific research questions and methodologies.). While the platform does enable searching of individual words and series of words (including using regular expressions) and provides rich, searchable metadata, users I have met at conferences sometimes express concern that the results may omit some hits or that they are not sure how to use the interface to produce searches as comprehensive as they would like. Downloading results for computational work is challenging for a basic user, and the words are not linked to an online dictionary as in the *Perseus Digital Library. Papyri.info* is a crown jewel of digital ancient studies, because it provides the features that it *does* very well. No one platform can do all things for all users.

Similarly, we see specialization (and thus different features) in other open-access VREs. The *Göttingen Old Testament Project*[9] produces digital editions of Coptic Christian Old Testament manuscripts utilizing the *Virtual Manuscript Room* environment created originally by the *Institut für Neutestamentliche Textforschung*, also using TEI markup (Behlmer 2017). The PAThs project in Rome has created an *archaeological atlas of Coptic literature*[10] by building an information hub about literary manuscript data – where codices were produced and found, where they are now archived or stored, where they have been published, which works are preserved on each codex, etc. (Buzi 2017; Buzi et al. 2018). The *Thesaurus Linguae Aegyptiae* (in collaboration with others) published an Egyptian Coptic lexicon formatted in TEI-XML, which CS then instrumentalized into an online dictionary[11], and the *Database and Dictionary of Greek Loanwords* in Coptic project subsequently contributed their Greek lemma list and definitions (Feder et al. 2018; Burns et al. 2019).

Collaborating with other projects or reusing their open-source data or technology also enables projects to excel in their own areas of research without having to reinvent the wheel in others. For the most part, digital papyrological projects collaborate with *Papyri.info* so that their data feeds into the common shared database. This allows for institutions with papyri collections to focus on their specific items while also contributing to a shared resource benefitting a wider scholarly community.

9 See https://www.uni-goettingen.de/en/digital+edition+des+koptischen+(sahidischen)+alten+Testaments/475974.html (Accessed: 25 June 2024).

10 See https://atlas.paths-erc.eu (Accessed: 25 June 2024).

11 See https://coptic-dictionary.org/about.cgi (Accessed: 25 June 2024).

The *Coptic Dictionary Online* (CDO) presents another example of reuse and collaboration among specialists. It contains the lexica from two projects, being the *Dictionary and Database of Greek Loan Words in Coptic* and the *Thesaurus Linguae Aegyptiae.* The CDO links each dictionary entry to individual words in the corpora published in CS's database; similarly, the CS database links back to the CDO word by word. In addition, entries for Egyptian Coptic words link to an online pdf of the most comprehensive print Coptic dictionary (by Crum [1939], hosted by yet another partner, the *Göttingen Old Testament project*). Entries for Greek loanwords link to the *Perseus* online Greek dictionary. The CS-team created and maintains the online interface that enables searching of the CDO and all of the interlinking of resources. Such a comprehensive, interlinked resource used widely internationally could not have been accomplished by one research unit alone.

Such accomplishments do not come without challenges, however. In Coptic, for example, Coptologists differ on what constitutes a word in the language. This may sound arcane, but the issue directly affects the creation of an online dictionary. Coptic is an agglutinative language, which means that different linguistic units (such as a subject pronoun and a verb) are bound together and are written together; additionally Coptic manuscripts are written in *scriptua continua,* with no white space between words or bound groups of words. Word segmentation is important for search, and also for creating lexical resources, such as a dictionary. Take the term for "idol-worshipper," *refšmšeeidolon.* Should we treat this term as one word with one lexical entry, since linguistically the whole item is a noun that takes one definite or indefinite article and as one term can be a subject of a verb? Or should we treat it as three words based on the morphemes that combine to create the term *(ref-šmše-eidolon)*? Where *šmše* means "to worship," *eidolon* is "idol", and *ref* is the prefix that indicates a term is a noun in the form of "the person who" does the thing that follows (the person who worships idols, or "idol-worshipper"). CS – with our interest in linguistics, part of speech annotation, and syntax annotation – treats the term as one word (a noun) with three morphemes. The TLA's research interests in creating their Egyptian Coptic lexicon concern (in part) tracing the Egyptian language through all its phases. Thus, it treats *ref-* as its own lexical entity as a lemma and gives it an entry in the *Coptic Dictionary Online* ("TLA lemma no. C3102"). Clicking on the link within that entry to find instances of the "word" *ref-* in the CS database, however, does *not* result in hits for all instances of *ref-* in our corpora, since we treat this morpheme as a prefix and not a lemma or word in and of itself; the query linking the CDO and CS corpora database is automated, so the different data models lead to a bit of a mismatch in a small number of cases (such as the morpheme *ref-*).

Deciding on a common definition of what constitutes a Coptic word or lemma before launching the CDO would have ground this collaboration to a halt. Instead, the projects agreed that some inconsistencies in mapping across our data were a small price to pay for the overall benefit of linking the dictionary to an online database of Coptic textual corpora. Sometimes these inconsistencies can be resolved in at least one direction; in the CS database, we do annotate a word like *refšmšeeidolon* as three morphemes, with a link for each morpheme to that morpheme's entry in the online dictionary. One might not be able to get to all the CS database hits for words that begin with *ref-* with one click from the CDO entry, but one can get them with a slight manual modification to the database query language. Additionally, one can access the dictionary entry for *ref-* with one click from the CS database. Manual mapping of entries alleviates some other inconsistencies, but such coding requires human labor, which can be challenging given the competitive and meagre funding opportunities for many humanities projects.

The very beginning of CS also benefitted from significant use of prior work, including open-source technology. Tito Orlandi's lexicon (published at CMCL) enabled us to create natural language processing tools that segmented Coptic text into words and tagged them with their parts of speech within the first year of the project. It easily cut a year off our initial work time. Instead of building our own database infrastructure, we adapted an open-source tool developed by linguists (including CS co-founder Zeldes) (Zeldes et al. 2009; Krause & Zeldes 2014). Again, this reuse allowed us to publish a searchable corpus of texts within months, not years. On the other hand, philologists and historians who are unfamiliar with corpus linguistics as a method can find the tool's search interface challenging. As a result, we have posted online tutorials and cheat-sheets to help users navigate the system and have invested development resources into modifying the tool for Coptic. While not perfect, the benefits of a robust, nearly out-of-the-box infrastructure outweigh the drawbacks, and certainly outweigh the costs of building an entirely new database infrastructure from scratch.

By necessity, I have not included all VREs for ancient studies or early Christian studies in this discussion of collaboration and reuse. Nonetheless, these examples illustrate some of the challenges resulting from specialization, disciplinary diversity, and intra-disciplinary methodological differences. Moreover, despite these challenges, open-source and open-access VREs administered by projects open to collaboration and data-sharing can stimulate much more robust research opportunities than more siloed endeavors.

3. Messy vs Clean Data

One interdisciplinary debate within Digital Humanities that directly affects VREs in ancient studies and early Christian studies is the degree to which we should clean our textual data. Philology as a discipline prizes accuracy and precision in text editions as well as in translations. Corpus linguists, computational linguists, and some digital humanists have a higher tolerance for mess.

"Messy" humanities data traditionally has been understood as large quantities of unstructured and unedited text (*big data*, Schöch 2013). Until recent years, scholars working in antiquity have not even had access to “big” ancient textual data in digital form. For Greek and Latin, *Perseus,* and *Open Philology* especially, but other projects as well, have contributed to large-scale digitization. For Coptic, Syriac, Ge’ez, and other languages, we are slowly moving towards what we might call *medium data.* Digital ancient studies experiences a push-and-pull between the desire for larger corpora of digital data we can search or analyze on the one hand, and the prioritizing of highly accurate, thoroughly peer reviewed editions on the other. In a 2013 conference paper about the creation and long-term viability of *Papyri.info,* Roger S. Bagnall cited the peer-review process as one of the factors slowing down the process of publishing more digital editions in their platform. Much of *Papyri.info* replicated in the digital realm – albeit in transformed ways – the scholarly form papyrologists were used to producing and using – the edition. And as such, it developed a peer review process before publishing editions online, much as print editions undergo peer review. The backlog of papyri or ostraca awaiting publication accumulated to the point at which the project board decided to publish editions that had not yet undergone the final round of peer review (Bagnall 2013). In digital publications, of course, we can release a new version with any corrections or editorial emendations quickly. In traditional print publications of editions and translations, scholars may labor for a decade or more ensuring the text is accurate with detailed apparatus notes, or commentary; except for extremely commonly read works, the appearance of revised editions or new editions by other scholars soon after the previous publication is rare. *Perseus* founder Gregory Crane commented on this phenomenon back in the 1980s in an early paper on Classics and “hypertext” (Crane 1987).

In a digital age, *messy* data might mean a variety of things. Inaccuracies in optical character recognition (OCR) when digitizing print editions. Typographical errors in transcriptions of ancient texts. Typographical errors in metadata. Misattribution of sources or inaccuracies in dating. For text with annotations for linguistic information such as part of speech, links to other resources, manuscript information, etc., any annotation errors also render data *messy.* Scholars editing, translating, and interpreting ancient texts often express that we are accustomed to working with highly accurate editions on all of these measures – accuracy of text, information about the work containing the text, translation, and so forth. The reality is that we find errors in print editions, as well. Our tolerance for mess, however, may be lower than what is required for working with automated digital methods. Accuracy rates of 98–99% for OCR, e.g., are considered quite high; in a million-word corpus, such a rate means ten to twenty thousand characters are affected – a number to which corpus linguists or computer scientists might be accustomed but that many philologists may find troubling (on OCR for historical languages generally, see Smith & Cordell 2018).

Some digital humanists have recently published work advocating for more tolerance for mess. Mess can involve *inaccuracies* in data or challenges to highly structured, formal systems and ideologies underpinning some computational work. In the latter understanding, as Losh et al (2016) have written, “‘Mess’ serves as a theoretical intervention in popular notions of digital media as neat, clean, and hyper-rational." Similarly, Katie Rawson and Trevor Muñoz argue that the debate over clean vs. messy data is an epistemological one: "The term 'cleaning' implies that a dataset begins as 'messy.' 'Messy' suggests an underlying order: it supposes things already have a rightful place, but they are not in it – like socks on the bedroom floor rather than in the bureau or the hamper" (Rawson & Muñoz 2019). In this view, cleaning a dataset – esp. normalizing, or annotating to create a structured data*set* out of unstructured *data* – involves imposing a pretheorized or presupposed order or model on the data. "The cleaning paradigm assumes an underlying, 'correct' order." Rawson & Muñoz (2019) advocate for embracing the diversity of unstructured data, and for allowing the querying and discovery of *uncleaned* data to point us to new understandings of the data and the communities that gave birth to it.

In philology – and here I specifically refer to ancient literature, especially biblical studies, rather than papyrology – the quest for *clean* textual data is connected to the quest for the *urtext. Clean,* here, is not perfectly spelled or accurately annotated text, but the earliest version of the work, the one closest to the original. Often the cleanest critical edition of a work matches no known manuscript 100 %. VREs and methodologies in manuscript studies take two different approaches to this pre-digital methodology. Tools and projects sometimes digitally replicate this traditional process, by transcribing (or creating VREs for transcribing) manuscript witnesses that will then be compared digitally to produce a critical edition. (Behlmer 2017; Huskey 2019) Tools such as *Juxta Commons* and *CollateX*[12] allow researchers to mark up parallel witnesses of the same text during the digital editing process (Wheeler & Jensen 2014).

Some digital humanists in Classics have also investigated how to produce digitally the print apparatus philologists are used to seeing; as a "data visualization," the apparatus is efficient and effective (Fischer 2019; Huskey 2022). Other projects, such as CS, publish digital editions of manuscript transcriptions (as well as earlier print editions) with metadata connecting versions of the same work to each other but without producing an apparatus or critical edition. In this way, at least, CS has embraced "mess." We certainly impose order on the text through our linguistic annotations, which employ a data model based in large part on the grammatical categories and syntax in Bentley Layton's *Coptic Grammar* – a work itself critiqued for aggressively creating and imposing new linguistic categories (Layton 2011; Shisha-Halevy 2006; Feder 2017). But with respect to the editions of Coptic literature, when publishing transcriptions of manuscripts, we transcribe the original text (however *messy*) and produce normalized and lemmatized text (the more *cleaned* textual data) as annotations on the original. Thus, the researcher may search for an expected, *cleanly* spelled word and also see all the instances of that term in its original spelling in our database. One can also pull up parallel manuscript witnesses, in the cases where we have published them. But we provide no critical edition or apparatus.

12 See https://collatex.net/about (Accessed: 25 June 2024).

4. Technical Standards

Traditionally in Digital Humanities, technical standards have provided three important functions. Standards lay the groundwork for how to mark up or process data so that subsequent projects do not have to reinvent the wheel. In this way, they provide a shared resource for humanists working in related research areas. To my mind, this is the most important aspect of digital standards – a community comes together to create a roadmap for each other and for researchers of the future. Even if not all aspects of set of standards fit every individual project in a field, the standards provide a starting point. Additionally, they will point other researchers toward known issues in digitization or computation in their scholarly field. For example, the PATHs project's data model includes more than one field for the author of a work – the "stated" author (as stated in the manuscript or work) and the "creator" (the verifiable historical author) (Buzi et al. 2018). Studying their data model and standards can help any project working on manuscripts and historical literature.

Standards also in theory help ensure consistency of data and annotations. For example, geographical locations annotated in the same way across a dataset allow researchers to query for a place and have a reasonable expectation of finding most if not all the instances of that location. Different textual data annotated according to the same standard by multiple projects also can be queried and analyzed in a comparative way. One such example is the *Universal Dependency dataset* (UD), in which corpora from over 100 languages have been annotated according to the same linguistic standards. Although Coptic has long been considered an *under-resourced* and perhaps even obscure language, its presence in the UD means that researchers have examined it alongside modern tongues, such as Danish and Chinese, to gain insight into language (Zeldes & Abrams 2018; Pinter et al. 2019; Chen et al. 2022).

Finally, this consistency in theory should lead to more interoperability between projects and research environments. Digital editions marked up according to a shared standard (such as TEI-XML) in one VRE should be publishable or editable in another VRE using the same standards. *Papyri.info* provides such an example; it aggregates papyri and ostraca digitized by multiple projects in one platform, a process possible in part due to shared use of the EpiDoc subset of the TEI-XML standards.

In practice, however, annotation is a process of interpretation. How to implement the same standard can vary. CS, the *Göttingen Old Testament* Project, and the *Canons of Apa John* project all have agreed to data share. We all use TEI-XML to annotate for manuscript information in our diplomatic transcriptions. However, the use some of the XML tags in slightly different ways, and we also have different understandings of what binds Coptic words into phrases called "bound groups." As a result, we created converter scripts to ensure true interoperability. These differences do not impose critical or unresolvable obstacles to collaborations, but they do point to the human element in data sharing. Additionally, interdisciplinary projects may find that no one set of standards can capture all the information their project will digitize and annotate. For example, CS releases our data in multiple formats according to different standards, because these standards have developed within specific fields for their individual disciplinary needs and research questions. While TEI-XML provides a robust tagset for digital editions, annotating for part of speech and syntax requires different kinds of annotations. Thus, our project releases our annotated corpora in multiple formats; each document is released as a *light* single TEI-XML file that captures manuscript information and some basic linguistic information (language of origin, lemma, part of speech), PAULA XML documents with full stand-off annotations for all aspects of our data model (including codicological and linguistic annotations), relational database files that contain full metadata and textual annotations used to populate the ANNIS database for querying of our corpora, and a SGML document with all annotations and metadata contained in one file.[13] We generate the files in these different formats from one master file. Moreover, we release the aforementioned UD corpus, which is a subset of our corpora with a high level of accuracy annotated according to the UD treebank syntax standards.

Communication and commitment to collaboration within disciplines and across disciplinary differences are just as important as technical standards. Such communication also extends beyond the scope of documentation. Documentation has long been raised as a key feature for sustainability and useability in Digital Humanities projects. It is also a common challenge, especially for projects running on limited funding and/or an abbreviated timeline for funding (Edmond & Morselli 2020). A project's standards as well as the decision-making process or technical investigations behind those standards can – and should – be documented in journal articles, project blogs, white papers, and "Read Me" files. Transparency about how a VRE functions, why it functions that way, and who contributed to the labor of the project is important (Keralis et al. 2023). In small research fields, successful projects cultivate a human mindset of collaboration and ongoing communication with users and research partners alongside providing documentation of standards.

5. Conclusion

Many discussions of VREs or other *tools* in Digital Humanities center on questions about sustainability (cf. the chapter by J. Apel in this volume, pp. 402–403). In building a tool or platform, project teams must consider the human labor required for creating it and supporting it over time, esp. as technology and standards change. VRE-teams need to consider how to provide sufficient training and documentation for users. Sustainability is also a human question as much as a technical one. Developing a VRE that is flexible enough to survive beyond initial startup funding (or to produce

13 See https://github.com/CopticScriptorium/corpora (Accessed: 25 June 2024).

data in formats that survive) requires both technical expertise and personal commitments to such an approach. The topics I have addressed in this essay are embedded in conversations about Digital Humanities sustainability. Projects using VREs in Digital Humanities can benefit from considering how they might reuse existing data and tools – thus extending the lifecycle of other projects' output and possibly reducing the financial cost of the labor of development in their own projects. Conversations about technical standards and *messy* or *clean* data are essential when developing plans for sunsetting a project. Planning for collaboration at the outset can help projects avoid "reinventing the wheel" and also can enable use of their data or tools on a wider scale and longer timeline. Although a VRE is technical infrastructure, the questions, and methods essential to building and maintaining such a tool are deeply human.

References

Bagnall, R.S. (2013). Digital Presentation, Digital Editing, Digital Community. The Case of Papyrology. In *Meeting Abstracts. SBL Meeting 2013*. Baltimore: Society of Biblical Literature. URL: https://www.sbl-site.org/meetings/Congresses_Abstracts.aspx?MeetingId=23 (Accessed: 25 June 2024).

Behlmer, H. (2017). Die digitale Gesamtausgabe und Übersetzung des koptisch-sahidischen Alten Testaments. Ein neues Forschungsprojekt an der Akademie der Wissenschaften zu Göttingen, *Early Christianity,* 8(1), 97–107. DOI: https://doi.org/10.1628/186870317X14876711440169 (Accessed: 25 June 2024).

Burns, D.M., Feder, F., John, K., & Kupreyev, M. (2019). *Comprehensive Coptic Lexicon. Including Loanwords from Ancient Greek* [data set]. DOI: https://doi.org/10.17169/REFUBIUM-2333 (Accessed: 25 June 2024).

Buzi, P. (2017). Tracking Papyrus and Parchment Paths. An Archaeological Atlas of Coptic Literature. Literary Texts in Their Geographical Context. Production, Copying, Usage, Dissemination and Storage (PAThs), *Early Christianity,* 8(4), 507–516. DOI: https://doi.org/10.1628/186870317X15100584934630 (Accessed: 25 June 2024).

Ead., Bogdani, J., & Berno, F. (2018). The 'PAThs'-Projekt. An Effort to Represent the Physical Dimension of Coptic Literary Production (Third-Eleventh Centuries), *Comparative Oriental Manuscript Studies Bulletin,* 4(1), 39–58. DOI: https://doi.org/10.25592/uhhfdm.253 (Accessed: 25 June 2024).

Chen, X., Gerdes, K., Kahane, S., & Courtin, M. (2022). The Co-Effect of Menzerath-Altmann Law and Heavy Constituent Shift in Natural Languages. In M. Yamazaki, H. Sanada, R. Köhler, Sh. Embleton, R. Vulanović & E.S. Wheeler (Eds.), *The Co-Effect of Menzerath-Altmann Law and Heavy Constituent Shift in Natural Languages* (pp. 11–24). Mouton: De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110763560-002 (Accessed: 25 June 2024).

Crane, G. (1987). From the Old to the New. Intergrating Hypertext into Traditional Scholarship. In *Proceedings of the ACM Conference on Hypertext* (pp. 51–55). New York: Association on Computing Machinery. DOI: https://doi.org/10.1145/317426.317432 (Accessed: 25 June 2024).

Id. (1998). The Perseus Project and Beyond. How Building a Digital Library Challenges the Humanities and Technology, *D-Lib Magazine*, no pag. URL: http://www.dlib.org/dlib/january98/01crane.html (Accessed: 25 June 2024).

Crum, W. E (1939). *Ein koptisches Wörterbuch.* Oxford: Clarendon Press.

Depauw, M., & Gheldof, T. (2013). Trismegistos. An Interdisciplinary Platform for Ancient World Texts and Related Information. In Ł. Bolikowski, V. Casarosa, P. Goodale, N. Houssos, P. Manghi, & J. Schirrwagen (Eds.), *Theory and Practice of Digital Libraries. TPDL 2013. Selected Workshops.* Cham: Springer [= *Communications in Computer and Information Science,* 416]. DOI: https://doi.org/10.1007/978-3-319-08425-1_5 (Accessed: 25 June 2024).

Edmond, J., & Morselli, F. (2020). Sustainability of Digital Humanities Projects as a Publication and Documentation Challenge, *Zeitschrift für Dokumentation,* 76, 1019–1031.

Feder, F. (2017). Review of *A Coptic Grammar* by Bentley Layton, *Orientalistische Literaturzeitung,* 112(2), 108–12. DOI: https://doi.org/10.1515/olzg-2017-0035 (Accessed: 25 June 2024).

Id., Kupreyev, M., Manning, E., Schroeder, C.T., & Zeldes, A. (2018). A Linked Coptic Dictionary Online. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 12–21). Santa Fe, New Mexico: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W18-4502 (Accessed: 25 June 2024).

Fischer, F. (2019). Digital Classical Philology and the Critical Apparatus. In M. Berti (Ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution* (pp. 203–220). Berlin/Boston: De Gruyter Saur. DOI: https://doi.org/10.1515/9783110599572-012 (Accessed: 25 June 2024).

Forstall, Ch.W., Jacobson, S.L., & Scheirer, W.J. (2011). Evidence of Intertextuality. Investigating Paul the Deacon's *Angustae Vitae, Literary and Linguistic Computing,* 26(3), 285–296. DOI: https://doi.org/10.1093/llc/fqr029 (Accessed: 25 June 2024).

Huskey, S. (2019). The Digital Latin Library. Cataloging and Publishing Critical Editions of Latin Texts. In M. Berti (Ed.), *Digital Classical Philology* (pp. 19–34). Berlin/Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110599572-003 (Accessed: 25 June 2024).

Id. (2022). The Visual [Re]Presentation of Textual Data in Traditional and Digital Critical Editions, *Magazén,* 1. DOI: https://doi.org/10.30687/mag/2724-3923/2022/05/005 (Accessed: 25 June 2024).

Keralis, S.D.C., Mirza, R., & Seale, M. (2023). Librarians' Illegible Labor. Toward a Documentary Practice of Digital Humanities. In M.K. Gold & L.F. Klein (Eds.), *Debates in the Digital Humanities* 2023 (no pag.). Minneapolis: University of Minnesota Press. URL: https://dhdebates.gc.cuny.edu/read/debates-in-the-digital-humanities-2023/section/c8bfbcfa-1500-41c2-a1d7-63b8c81b627f#ch20 (Accessed: 25 June 2024).

Krause, Th., & Zeldes, A. (2014). ANNIS3. A New Architecture for Generic Corpus Query and Visualization, *Digital Scholarship in the Humanities,* 31(1), 118–139. DOI: https://doi.org/10.1093/llc/fqu057 (Accessed: 25 June 2024).

Layton, B. (2011). *A Coptic Grammar.* 3. ed. Wiesbaden: Harrassowitz [= *Porta Linguarum Orientalium. Neue Serie,* 20].

Losh, E., Wernimont, J., Wexler, L., & Wu, H.-A. (2016). Putting the Human Back into the Digital Humanities. Feminism, Generosity, and Mess. In M. K. Gold & L. F. Klein (Eds.), *Debates in the Digital Humanities* 2016 (no pag.). Minneapolis: University of Minnesota Press. URL: https://dhdebates.gc.cuny.edu/read/untitled/section/cfe1b125-6917-4095-9d56-20487aa0b867#ch10 (Accessed: 25 June 2024).

Okuda, N., Kinnison, J., Burns, P., Coffee, N., & Scheirer, W. (2022). Tesserae Intertext Service, *Digital Humanities Quarterly,* 16(1), 1–61. URL: http://www.digitalhumanities.org/dhq/vol/16/1/000602/000602.html (Accessed: 25 June 2024).

Orlandi, T. (2021). Reflections on the Development of Digital Humanities, *Digital Scholarship in the Humanities,* 36(2), 222–229. DOI: https://doi.org/10.1093/llc/fqaa048 (Accessed: 25 June 2024).

Pinter, Y., Marone, M., & Eisenstein, J. (2019). Character Eyes. Seeing Language through Character-Level Taggers. In T. Linzen, G. Chrupala, Y. Belinkov & D. Hupkes (Eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP. Analyzing and Interpreting Neural Networks for NLP* (pp. 95–102). Florenz: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/W19-4811 (Accessed: 25 June 2024).

Rawson, K., & Muñoz, T. (2019). Against Cleaning. In M. K. Gold & L. F. Klein (Eds.), *Debates in the Digital Humanities* 2019 (no pag.). URL: https://dhdebates.gc.cuny.edu/read/untitled-f2acf72c-a469-49d8-be35-67f9ac1e3a60/section/07154de9-4903-428e-9c61-7a92a6f22e51#ch23 (Accessed: 25 June 2024).

Schöch, Ch. (2013) Big? Smart? Clean? Messy? Data in the Humanities, *Journal of Digital Humanities,* 2(3), no pag. URL: http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities (Accessed: 25 June 2024).

Schroeder, C. T. (2019). Cultural Heritage Preservation and Canon Formation. What Syriac and Coptic Can Teach Us about the Historiography of the Digital Humanities. In G. Frank, S. Holman & A. Jacobs (Eds.), *The Garb of Being. Embodiment and the Pursuit of Holiness in Late Ancient Christianity* (pp. 318–345). New York: Fordham University Press.

Ead., & Zeldes, A. (2020). A Collaborative Ecosystem for Digital Coptic Studies, *Journal of Data Mining & Digital Humanities,* 1–9. [= *Numéro spécial sur la collecte, la préservation et la diffusion du patrimoine culturel menacé pour de* *nouvelles compréhensions grâce à des approches multilingues*]. DOI: https://doi.org/10.46298/jdmdh.5969 (Accessed: 25 June 2024).

Eid. (2016). Raiders of the Lost Corpus, *Digital Humanities Quarterly,* 10(2), 1–38. URL: http://digitalhumanities.org/dhq/vol/10/2/000247/000247.html (Accessed: 25 June 2024).

Shisha-Halevy, A. (2006). Review of *Coptic Grammar. 2. ed.* by Bentley Layton, *Orientalia,* 75(1), 132–133. URL: https://arielshishahalevy.huji.ac.il/publications2006c (Accessed: 25 June 2024).

Smith, D. A., & Cordell, R. (2018). *A Research Agenda for Historical and Multilingual Optical Character Recognition.* URL: http://hdl.handle.net/2047/D20297452 (Accessed: 25 June 2024).

Wheeler, D., & Jensen, K. (2014). Juxta Commons [poster], *Journal of Digital Humanities,* 3(1), no pag. URL: https://journalofdigitalhumanities.org/3-1/juxta-commons (Accessed: 25 June 2024).

Zeldes, A., & Abrams, M. (2018). The Coptic Universal Dependency Treebank. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (pp. 192–201). Brüssel: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/W18-6022 (Accessed: 25 June 2024).

Zeldes, A., Ritz, J., Lüdeling, A., & Chiarcos, Ch. (2009). ANNIS. A Search Tool for Multi-Layer Annotated Corpora. In *Proceedings of Corpus Linguistics 2009.* Liverpool: American Association of Corpus Linguistics. URL: http://ucrel.lancs.ac.uk/publications/cl2009 (Accessed: 25 June 2024).

Research Data Management

Jochen Apel

https://orcid.org/0000-0002-0395-4120

Abstract This chapter introduces the area of research data management. Based on the FAIR Data Principles, the chapter outlines the main features of various aspects and aims of data management – from project planning and data organization in the project to the publication and preservation of research data.*

Keywords Research Data, Research Data Management, FAIR principles, Humanities, Digital Humanities

1. Research Data and Research Data Management

Structured and planned handling of research data is a central requirement for every research project, of course also in the field of theology. With the increasing use of digital tools and methods, the requirements for effective and sustainable data management that ensures the quality, traceability, and reusability of research results are growing. This chapter explains the basic principles of research data management (RDM). The term research data is understood in the following way:

> Research data are digital data generated, collected, or compiled in the research process and on the basis of which scientific hypotheses, models, or theories are formed and confirmed or refuted.[1]

According to this definition, digital data become research data through the specific *epistemic roles* they play in the research process. This seems particularly relevant for research in the humanities, where, in many cases, data is not generated as part of a research project, but is already available, *only becoming research data* through scien-

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson, and revised by Kevin Wunsch.

1 The author of this chapter first formulated proposed definition for a local website in 2014. Cf. https://web.archive.org/web/20230320185206/https://data.uni-heidelberg.de/faq.html (Accessed: 19 June 2024).

Jochen Apel: Research Data Management, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 393–405. DOI: https://doi.org/10.11588/heibooks.1521.c21961

tific inquiry. If source material becomes the subject of research, the corresponding digital data becomes research data. The proposed definition is purposefully relatively broad. It can encompass the following types of data: digital sources and digitized versions of sources (text, image, video, and video files or 3D data), edits of source (e.g., an OCR-generated text or an edition encoded in TEI-XML), results of analyses (e.g., statistical results of a quantitative analysis), or databases in which information is structurally compiled.

The relevant literature shows many attempts to define research data. These attempts are summarized by Geiger (2023). The landscape of research data in the humanities is a heterogeneous field in which different data types, formats, and structures are relevant and in which there are only established standards to a certain extent or in certain sub-areas (cf. Pempe 2012).

Thus, one can more precisely define what is meant by RDM. Two definitions from the relevant literature are worth mentioning, which together span the spectrum of meaning of the term:

> Research data management refers to all measures ensuring digital research data is usable. However, the requirements for doing so vary depending on the different purposes for which research data is to be used. Four types of purpose can be distinguished:
>
> - Use as a working copy for research;
> - the subsequent use of research data for later research;
> - storage as documentation of sound academic practice; and
> - storage to comply with legal or other non-research requirements (Enke & Ludwig 2013, 13).

> Research data management concerns the organization of data, from its entry to the research cycle through to the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information (Whyte & Tedds 2011).

While the first definition particularly identifies the usability of data as the goal of RDM, the essential aspect of the data lifecycle image used in the second definition is the emphasis on the integration of the different phases of the research process. The handling of research data in later phases of a project depends on setting the agenda in the earlier phases. For example, if one works with data to which third parties hold rights and wants to publish this data, one must clarify the necessary rights already during the process of data collection. Assigning resources to the comprehensive documentation of the research data is imperative, if the data should be re-used by future projects.

2. FAIR Data Principles and the Aims of Research Data Management

The previous section, following Enke and Ludwig, argued that RDM is concerned with keeping research data usable. However, what does usability mean in detail? What characteristics must research data have to be usable? Wilkinson et al. (2016) attempt to answer these questions by providing the FAIR Data Principles.

FAIR stands for the four elements: *Findable, Accessible, Interoperable and Reusable.* These concepts are detailed out as follows:

> Findable
> F1. (Meta)data are assigned a globally unique and persistent identifier.
> F2. Data are described with rich metadata (see R1 below).
> F3. Metadata clearly and explicitly contain the identifier of the data they describe.
> F4. (Meta)data are registered or indexed in a searchable resource.
>
> Accessible
> A1. (Meta)data are retrievable by their identifier using a standardized communications protocol.
> A1.1 The protocol is open, free and universally implementable.
> A1.2 The protocol allows for authentication and authorization procedure, where necessary.
> A2. Metadata are accessible, even if the data are no longer available.
>
> Interoperable
> I1. (Meta)data use a formal, accessible, shared and widely applicable language for knowledge representation.
> I2. (Meta)data contain vocabularies that follow FAIR principles.
> I3. (Meta)data contain qualified references to other (meta)data.
>
> Reusable
> R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.
> R1.1. (Meta)data are released with a clear and accessible data usage license.
> R1.2. (Meta)data are associated with detailed provenance.
> R1.3. (Meta)data meet domain-relevant community standards.[2]

Without discussing the various aspects of the FAIR principles in detail, the following key characteristics can be identified. *Findability* is based on comprehensive documentation and description using metadata and the use of persistent digital identifiers

2 See https://www.go-fair.org/fair-principles (Accessed: 19 June 2024).

(e.g., *Digital Object Identifier* (DOI)), which form the basis for stable referencing in publications, reference systems, and search engines. *Accessibility* is about the attempt to make data available as openly as possible and as protected as necessary. Ideally, research data should be published as Open Research Data, under the most permissive license possible. Should doing so not be possible, however, the data can be made available to authorized users using suitable authentication methods. If this is also not possible, at least the descriptive metadata should be publicly available. *Interoperability,* in turn, is essentially based on standards for data and metadata. The use of standardized data structures, non-proprietary data formats, or standardized vocabularies are the basis for the easy usability of the data. If data is available in machine-readable form, its reuse can be automated as well as the possible integration of the data with other data sets. *Reusability* is based on the rich description and documentation of the content, including provenance information, as well as on legal stipulations as to how the data can be reused. This is accomplished by using suitable open content licenses (e.g., Creative Commons licenses).

Appropriate research infrastructures can support researchers in providing their data in accordance with the FAIR principles. While the description and documentation of the data must be carried out by the researchers themselves; the data will only be findable (in the sense of the FAIR principles) if the data repositories used to provide the data have suitable functionalities, e.g., support suitable metadata standards or offer persistent identifiers.

3. Frameworks and Guidelines for Data Management

In addition to the requirements for RDM arising from the research context, as well as those arising from the respective research practices, underlying conditions and guidelines must be maintained. Researchers should consider these underlying conditions when planning a project. Third-party funding bodies now universally formulate requirements for the RDM of supported projects. These bodies do so either in the form of central guidelines, such as those issued by the DFG and the EU, and/or in the form of specific prerequisites within the individual program lines, as is the case with the BMBF.[3]

However, not only third-party funders but also universities and other research institutions formulate rules for handling research data.[4] In some cases, this is done in

3 See https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/forschungsdaten/index.html; https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_de.pdf; https://www.openaire.eu/rdm-in-horizon-europe-proposals; and https://forschungsdaten.info/themen/informieren-und-planen/foerderrichtlinien. All addresses were accessed on 19 June 2024.

4 Cf. https://www.forschungsdaten.org/index.php/Data_Policies (Accessed: 19 June 2024).

dedicated data policies or in codes of conduct on data protection in order to safeguard good academic practice. In addition, the scientific communities themselves – often by their respective learned societies – as well as publishers and academic journals formulate frameworks for handling research data.[5]

4. Data Management Plans

A data management plan can serve as starting point and basis for RDM. A data management plan is a document that sets out which data is collected or used and how this data is to be handled in the project as well as afterwards. Ideally, the data management plans should be regularly consulted and updated as needed throughout the project.

The purpose of a data management plan is to make well-founded decisions on how to handle the research data of one's own project, to identify risks and challenges at an early stage, to establish uniform procedures and standards for the joint use of data in cooperative projects, and thereby ensure the sustainability of the data, saving time and effort by designing efficient processes. William Michener has formulated the central topics that a data management plan should address in the form of a checklist with ten aspects to consider (Michener 2015)[6]:

1. Determine the research sponsor requirements.
2. Identify the data to be collected.
3. Define how the data will be organized.
4. Explain how the data will be documented.
5. Describe how data quality will be assured.
6. Present a sound data storage and preservation strategy.
7. Define the project's data policies.
8. Describe how the data will be disseminated.
9. Assign roles and responsibilities.
10. Prepare a realistic budget.

There are a variety of templates, checklists, and web-based tools for creating a data management plan for a project (Dierkes 2021, 310). Tools such as the RDMO and DMPonline services are particularly helpful, as they use comprehensive question-

5 Cf. https://forschungsdaten.info/themen/ethik-und-gute-wissenschaftliche-praxis/leitlinien-und-policies (Accessed: 19 June 2024).

6 See also https://forschungsdaten.info/themen/informieren-und-planen/datenmanagementplan and the Science Europe guide: https://www.scienceeurope.org/media/4brkxxe5/se_rdm_practical_guide_extended_final.pdf. Both addresses were accessed on 19 June 2024. In addition to the list below, they also address issues such as data reuse, the legal and ethical framework and data sharing.

naires to draw attention to all potentially relevant aspects to be considered in data management planning.

5. Data Management in the Project (hot data)

Key aspects of data management concern the planned handling of research data over the course of a project. In many respects, these aspects are specific to the subject, method, and data specific and thus cannot be adequately explained in an introductory overview. However, they are addressed in other chapters of this handbook, in which concrete case studies and specific methods are discussed. This subsection focuses on generic aspects relevant to any research project.

5.1 Data Collection

The way in which data is collected for a research project depends on the project. In humanities projects, the data is often not collected by the researchers themselves, but provided by third parties, e.g., as publicly available digital copies from libraries and archives. This information is important not only because of the required attribution information or possible requirements for use and transfer of the data of those providing the data, but also because questions regarding the preservation and publication of the data may be answered differently. If the data on which a project is based has been published by a trustworthy provider, such as a library, then this institution will also ensure the long-term preservation of the data. Thus, the project does not need to find a solution for these aspects and can focus primarily on dealing with working copies, which are only needed during the project. Of course, when using third-party data, it may also be the case that the project generates its own research data derived from original data sets, e.g., annotations, tabular evaluations, or statistical analyses; the long-term preservation and publication of this data is thus subject of the project's own research data management.

Regardless of how data is collected, the documentation required for the long-term usability of the data should be carried out as early as possible, ideally using relevant metadata standards. The documentation can take various forms, which can also be combined, for example in Read Me files, structured metadata databases, a wiki, or directly in the data management plan.[7]

7 See https://forschungsdaten.info/themen/beschreiben-und-dokumentieren/datendokumentation (Accessed: 19 June 2024).

5.2 Storage, Backup, and Deletion

For every research project, the question arises as to where and how the project-relevant research data is stored, such as on the hard disk of a local PC, on an institutional server, a university's central storage, or commercial provider. The use of central services is particularly recommended for large amounts of data. Every research project requires a suitable backup strategy. The so-called 3-2-1 rule serves as a rough guide here: three copies of data on two different storage media, one of which is at an external location (see Krogh 2009, Chapter 6).

If the data is stored on a personal computer, the owners are required to set up their own backup. If the institutional servers or other service providers are used, the provision of storage and backup solution falls within the scope of these providers. In principle, these solutions should be prefered. The latter solutions should therefore be given preference in principle. For small data volumes, the sync-and-share services now offered relatively widely by research institutions offer a low threshold solution.

In addition, research data management may already include selection processes regarding the deletion of files at this stage. Raw data on which the research is based should never be deleted during the project. Processed versions of the data can be deleted if they are no longer required. However, each individual step should always be documented to enable retracing the corresponding processing steps on the basis of the raw data if necessary. Processed versions of the data should be retained if they are essential building blocks of a publication. When the research project reaches its end, additional regulations may need to be made (e.g., regarding the deletion period for certain data).

5.3 Data Sharing

When choosing a suitable storage system, in addition to data security, backup, and costs, it is relevant whether the data in the research project is to be shared with other people and in which form this should take place. Will the data be shared within a research group or a project network? Does this require parallel access to and/or simultaneous work in files? Should cooperation partners have access to some or all the data on an ad hoc or permanent basis? Here, it is important to ensure that the selected services ideally already provide suitable rights management that can be adjusted as finely grained as possible and in which the corresponding access and editing permissions can be managed.

5.4 Data Organization

Like the aspects discussed above, data organization depends heavily on the respective research project and the relevant data material: image data must be handled differently than text corpora or data in tabular form; researchers working alone have different requirements than cooperative research projects in which a shared database is used; and large volumes of data must be handled differently than small volumes. Nevertheless, there are some generic recommendations for data organization.

Conventions on folder structure and file naming form the basis for a sensible form of data organization. The handbook *The Turing Way Community* (2022, chapter "Research Data Management") provides a straightforward introduction. Research data is usually stored in a folder structure. The naming conventions for folders should follow a systematic approach, such as chronological sorting, sorting according to the survey methods used, or assignment to individual sub-projects. Within the folders, the files should be named in a systematic way like using the creation date in the form YYYYMMDD as their respective names. Data types, data collection method, the name of the researcher or initials (especially in cooperative projects), or version numbers might be additional elements of the naming scheme.

The guiding principle should be that the file names should provide context for the respective file to distinguish it from similar files and from other versions of the same file. In addition, one might also consider suitable software tools to support data organization. Text-based data can be efficiently managed and versioned in a Git system. File naming tools can help name large quantities of files according to a uniform scheme.

5.5 Choice of Data and File Formats

Regarding the preparation of data in line with the FAIR principles, attention should also be given early on to the selection of suitable data and file formats. It is particularly important to differentiate between proprietary and non-proprietary (Open Source) formats as well as between binary and text formats.

When possible, non-proprietary formats should be chosen. However, a distinction must be made between binary and text formats. Binary formats are generally less memory-intensive, and many software products process and deliver binary formats. Nevertheless, with respect to the long-term preservation of data, one can (also) store their data in a textual format (if a corresponding migration is possible). Different formats exist for different means and for different purposes, namely, active work with the data in the project and later archiving and preservation. Wether data migration of the format is possible without loss of functionality and into any repository or archival software should be evaluated early on in the project phase. If conversion is only possible with a loss of information, it is important to consider whether this loss affects significant properties or not. For example, a table in Excel format may contain formatting such as bold column headings that are lost when converted to CSV format. However, if the formatting is not a significant feature for archiving, then a corresponding loss of information due to the migration can be acceptable. The Swiss Coordination Office for the Permanent Archiving of Electronic Records (KOST) provides a helpful overview of more than 50 common data formats and their suitability for long-term preservation.[8]

6. Archiving and Publication (cold data)

6.1 Repositories and Data Publication

By the end of a project, research data should be handed over to a trustworthy research data repository or data center, which takes over the sustainable long-term provision of the research data.

The following questions serve as a guide when selecting a suitable publication location for research data: Are there any special protection requirements for making the data available? Should the data be kept inaccessible to the public and only available upon request and under certain conditions? In this case, a service or repository must be found that can guarantee this form of *closed access.* If not, the next step should be to check whether there are suitable specialist repositories. These are usually the most suitable publication venues for research data due to their specialization in certain disciplines and/or data types. Data publications are visible in the subject-specific context in a joint collection with other relevant data from the subject. Subject repositories also support established metadata standards in the subject, offer specific search or visualization functionalities, and the operators of a subject repository may also be able to guarantee comprehensive curation and review of the data on the basis of relevant subject expertise.

Particularly if no suitable subject-specific service is available, institutional repositories or other generic, i.e., non-subject-specific services can be used. These do not offer the subject-specific functionalities described above but are nevertheless trustworthy publication venues that also comply with the FAIR principles. In particular, the strong connection to local research data support units and the benefit of direct on-site support might be strong arguments in favor of using institutional structures.

Unless there are specific reasons to the contrary, the data should be published as open research data, whereby open content licenses should be used wherever possible to ensure the broadest possible and lowest-threshold reuse of the data. Similar to Open Access publications, Creative Commons licenses have been established in

8 See https://kost-ceco.ch/cms/kad_intro_de.html (Accessed: 19 June 2024).

the field of research data as the *de facto* standard. This standard is questioned by researchers from big data fields, such as automated text analysis, who argue that research data should be placed in the public domain as far as possible, for example, by using a CC0 waiver (Brettschneider et al. 2021).

The international directory of research data repositories, *re3data*, lists a total of 22 subject repositories from the field of theology in May 2023.[9] However, depending on the field of research, other intellectual repositories in the humanities may also be suitable publication venues, e.g., the existing offerings and those currently under development of the humanities NFDI consortia NFDI4Culture, NFDI4Memory, NFDI4Objects, and Text+,[10] or the repositories of the CLARIAH network.[11] The data centers working group in the DHd – *Association for Digital Humanities in the German Speaking Areas* – also offers assistance in the search for suitable platforms.[12] Individual websites or web-based databases, which are a frequent result of humanities projects as "presentation layers" of research data, are a special case. The choice of these individual presentation formats may often make sense due to the heterogeneity of research questions in the humanities; however, the problem of sustainability arises immediately: how and by whom should such data products be able to be operated in the long term beyond the duration of the project? Sustainability is only possible if an infrastructure partner is involved at an early stage and should include the fallback option to turn off the web presentation while keeping the raw data available through a repository.

6.2 Long-Term Archiving

The problem of long-term digital preservation is particularly pressing for relevant research data in the humanities that remains relevant over very long periods of time. Digital long-term archiving has three aspects (cf. Liegmann et al. 2010):

- Bitstream Preservation
- Preservation of Functionality
- Preservation of Usability

Digital preservation cannot be ensured by individual researchers or research projects; rather, it requires technically and organizationally elaborate, sustainable infrastructures carrying out this overarching task. Nevertheless, researchers can con-

9 See https://www.re3data.org/search?subjects[]=107%20Theology (Accessed: 19 June 2024).

10 See https://nfdi4culture.de, https://4memory.de, https://www.nfdi4objects.net and https://www.text-plus.org. All addresses were accessed on 19 June 2024.

11 See https://www.clariah.de/publizieren-archivieren (Accessed: 19 June 2024).

12 See https://dhd-ag-datenzentren.github.io (Accessed: 19 June 2024).

tribute directly to the preservability of their generated research data, e.g., by using open, non-proprietary data formats or converting their data into such formats where possible. Doing so enables the maintenance of the functionality of the data in the long term. These formats are more likely to be supported in the long run and operators of preservation services are more likely to be able to transfer the data to new formats if the existing formats are no longer supported. Through adequate documentation and the provision of rich metadata, researchers also help to ensure that the data remains not only functional, but also usable, since only in this way can the data be understood, adequately interpreted, and contextualized.

7. Conclusion

Research data management is, without a doubt, a crucial element of the research process. The Digital Humanities (like other fields of research) can only fully develop the potential with high-quality research data organized according to the FAIR principles. This crucial role is summarized by Alma Gold as follows:

> Data is the currency of science, even if publications are still the currency of tenure. To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself" (Gold 2007).

RDM is essentially an activity of the researchers themselves, thus an intrinsic part of the research process (cf. also Lemaire 2018, 245). However, there are broad and diverse range of services that support researchers by providing advice and the necessary infrastructure and tools. Such services include the NFDI consortia, the CLARIAH network, and the DHd, but also institutional service centers for RDM remain key contact points. RDM that is consistently designed to implement the FAIR Data Principles improves the quality of research and its results and is essential for the seamless execution of a research project as well as for follow-up research based on the current project. Plainly stated, there is no digital research without RDM. You might manage your data efficiently or inefficiently, more FAIR, or less FAIR, but there never is no RDM.

References

Brettschneider, P., Axtmann, A., Böker, E., & Suchodoletz, D. v. (2021). Offene Lizenzen für Forschungsdaten, *o-bib. Das offene Bibliotheksjournal,* 8(3), 1–22. DOI: https://doi.org/10.5282/O-BIB/5749 (Accessed: 19 June 2024).

Dierkes, J. (2021). Planung, Beschreibung und Dokumentation von Forschungsdaten. In M. Putnings, H. Neuroth, & J. Neumann (Eds.), *Praxishandbuch Forschungsdatenmanagement* (pp. 303–326). Berlin/Boston: De Gruyter Saur. DOI: https://doi.org/10.1515/9783110657807-018 (Accessed: 19 June 2024).

Enke, H., & Ludwig, J. (Eds.). (2013). *Leitfaden zum Forschungsdaten-Management.* Boizenburg: Verlag Werner Hülsbusch. URL: https://www.forschungsdaten.org/images/b/b0/Leitfaden_Data-Management-WissGrid.pdf (Accessed: 19 June 2024).

Geiger, J. D. (2023). Daten/Forschungsdaten. In AG Digital Humanities Theorie des Verbandes Digital Humanities im deutschsprachigen Raum e. V. (Ed.), *Begriffe der Digital Humanities. Ein diskursives Glossar.* Wolfenbüttel: Herzog August Bibliothek [= *Zeitschrift für digitale Geisteswissenschaften. Working Papers,* 2]. DOI: https://doi.org/10.17175/WP_2023_003 (Accessed: 19 June 2024).

Gold, A. (2007). Cyberinfrastructure, Data, and Libraries, 1, *D-Lib Magazine,* 23(1/2), no pag. DOI: https://doi.org/10.1045/september20september-gold-pt1 (Accessed: 19 June 2024).

Kailus, A. (2023). Handreichung für ein FAIRes Management kulturwissenschaftlicher Forschungsdaten. V. 1.0.3. URL: https://nfdi4culture.de/go/E3625 (Accessed: 19 June 2024).

Krogh, P. (2009). *The DAM Book. Digital Asset Managament for Photographers.* 2. ed. Sebastopol: O'Reilly Media.

Lemaire, M. (2018). Vereinbarkeit von Forschungsprozess und Datenmanagement in den Geisteswissenschaften. Forschungsdatenmanagement nüchtern betrachtet, *o-bib. Das offene Bibliotheksjournal,* 5(4), 237–247. DOI: https://doi.org/10.5282/O-BIB/2018H4S237-247 (Accessed: 19 June 2024).

Liegmann, H., & Neuroth, H. (2010). Einleitung. In H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, & K. Huth. (Eds.), *nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung.* Version 2(3) (pp. 1–10). Boizenburg: Verlag Werner Hülsbusch. URL: https://nbn-resolving.de/urn:nbn:de:0008-2010071949 (Accessed: 19 June 2024).

Michener, W. K. (2015). Ten Simple Rules for Creating a Good Data Management Plan, *PLOS Computational Biology,* 11(10), 1–9. DOI: https://doi.org/10.1371/journal.pcbi.1004525 (Accessed: 19 June 2024).

Pempe, W. (2012). Geisteswissenschaften. In N. Heike, S. Strathmann, A. Oßwald, R. Scheffel, J. Klump, & J. Ludwig (Eds.), *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme* (pp. 137–160). Boizenburg: Verlag Werner Hülsbusch.

The Turing Way Community. (2022). The Turing Way. A handbook for reproducible, ethical and collaborative research. Online: Zenodo. DOI: https://doi.org/10.5281/ZENODO.7625728 (Accessed: 19 June 2024).

Whyte, A., & Tedds, J. (2011). Making the Case for Research Data Management. In *DCC Briefing Papers.* Edinburgh: Digital Curation Centre. URL: https://www.dcc.ac.uk/guidance/briefing-papers/making-case-rdm (Accessed: 19 June 2024).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, Ph. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, Ch. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., C't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, Ph., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, Th., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship, *Scientific Data,* 3(1), 1–9. DOI: https://doi.org/10.1038/sdata.2016.18 (Accessed: 19 June 2024).

AI supported Text Production in the University

Johanna Gröpler[a], Margret Mundorf[b] and Nicolaus Wilder[c]

[a] https://orcid.org/0000-0002-1546-9843, [b] https://orcid.org/0000-0001-7992-6233, [c] https://orcid.org/0000-0002-5884-6768

Abstract Since the release of OpenAI's chatbot ChatGPT, tools based on artificial intelligence have also reached the wider public. This chapter discusses the opportunities and challenges for their use at the university level.*

Keywords Artificial Intelligence, Text Production, University Education

1. AI and the Disruption of the Academic World

The public has previously only associated *artificial intelligence* (AI), connected with natural language, with service chatbots that provide information on request, such as the opening hours of a library or the general terms and conditions of a telephone company. These were previously recognizable by their limited "knowledge," understood as information processing skills tailored to a specific area of application (Adiwardana & Luong 2020). This limited perspective changed abruptly in November 2022 when the US company OpenAI released ChatGPT, an AI-based tool for conversations. The application enables one to conduct human-like dialogs with an algorithm-based system in a chat window using *prompts* in natural language, such as asking questions, have facts explained, or perform more complex actions. As the system's output appears plausible and coherent, many users do not realise that they are not based on indisputable data, but are merely statistical calculations. The chats with the free version of ChatGPT[1] based on the GPT-3 language model and the improved version GPT-3.5 were in many ways indistinguishable from interaction with a human.

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

1 On 14 March 2023, the successor model, GPT-4, was released as a paid version and expanded to include multimodal capabilities and access to data from the internet. Since November 2023, it has

Johanna Gröpler, Margret Mundorf and Nicolaus Wilder: AI supported Text Production in the University, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 407–427. DOI: https://doi.org/10.11588/heibooks.1521.c21962

The Turing test – developed to distinguish machine thinking from human thinking – seemed to have been passed (Borchers 2022).

In 2019, Springer Nature published the first specialist publication on lithium-ion battery research generated entirely by AI under the pseudonym *Beta Writer* (2019). The research summarized and referenced all articles on the topic of lithium batteries, which provided an outlook on how the literature review step could be automated in scientific work in the future (Wilder et al. 2022, 217f.). In education, the hype surrounding ChatGPT triggered dystopian panic, utopian-naïve glorification, and heated debates on written examination forms such as term papers and theses: How can a text document a person's cognitive process and level of knowledge when the work could have possibly been generated by a machine? While there were skeptical voices saying that the door was open to attempts at cheating, which could now increase dramatically, others, including the author's marketing departments, emphasized the advantages of using AI in education, which should spark nothing less than an educational revolution: every person, regardless of their learning level, their abilities, their socio-economic conditions, etc., would now have an individual learning companion at their disposal to support the learning process in all phases and thus make a significant contribution to the acquisition of skills. The controversial discussion about the opportunities and risks of using AI-supported applications in an academic context therefore picked up speed immediately with the publication of ChatGPT and will continue to occupy and challenge the education sector. In essence, the controversial discourses that have arisen in the educational context are leading to the inevitable future task of fundamentally rethinking teaching, learning, and examination cultures (Weßels et al. 2022).

Disruptive effects are occuring on the academic writing process, ultimately on language and its use, as well as on the linguistic acquisition and transfer of knowledge. Further developments and their evaluation are difficult to predict and will be decided less by the technology of the AI systems themselves and more when used, how, and for what purpose they will be used (Limburg et al. 2023).

To outline and categorize the current state of the discussion, this chapter first explains how *Large Language Models* (LLMs) work, on which current tools like ChatGPT from *OpenAI*[2] or the open source project *Open Assistant*[3] are based. Generated outputs from these systems can thus be systematically analyzed along with the associated opportunities and risks, and future developments and their significance for the education sector can be assessed. Subsequently, the chapter presents examples of possible application scenarios in teaching and research and examines previous activities at universities to provide orientation for teachers and students. Challenges like data protection, copyright, and bias are also considered.

been possible to utilize specialized plugins for specific tasks or to create customized GPTs with one's own data. Since mid-May 2024, GPT-4o (omni) has been accessible at no cost to all users.

2 See https://chat.openai.com (Accessed: 18 June 2024).

3 See https://open-assistant.io/de (Accessed: 18 June 2024).

2. How AI Text Generators Work

The term *artificial intelligence* is a generic term for various systems with algorithms designed to perform human tasks independently. According to the current draft status of the EU AI Act (Art. 3), the draft law that provides for the regulation of AI systems within the European Union, AI systems are defined as

> software that is developed with one or more of the techniques and approaches listed in Annex I [these are machine learning, logic and knowledge-based concepts, and statistical approaches] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with software developed using one or more of the techniques and concepts listed in Annex (European Commission 2021, 46).

One area of particular importance for the performance of current systems is *Machine Learning* (ML), in which the system learns independently: patterns and regularities (a model) are extracted from training data (examples) based on a given algorithm to derive predictions (i.e., to apply the model to unknown examples). ML is generally used when the task to be performed is too complicated to be mapped in clear rules (Döbel et al. 2018, 8). Exemplary and typical areas of application for ML are image recognition or text generation. Determining the concept of "dog," for example, so precisely that it recognizes all dogs in images and does not identify any objects other than dogs is only possible with very detailed descriptions that cover all individual cases. In contrast, it is much more efficient to have an algorithm analyze a large data set of dog images. The algorithm can then develop a model of the concept of a "dog," which can then be used to determine further images with the distinction between dogs and non-dogs. The same applies to language, the essence of which cannot be reduced to syntactic rules.

In ML, a distinction is made between three different methods or learning styles: (a) *supervised,* (b) *unsupervised,* and (c) *reinforcement learning.* In *supervised learning* (a), the training data must first be provided with labels from humans in a complex process, an annotation with the assignment of an associated category for the desired output. For example, different animal photos are assigned the labels dog, cat, or mouse – which must already be known as a result of this learning style. The machine then essentially creates its model using the relationship between label, image, and text. Since the predictions of the model can be compared with the known correct results, the learning process can be "monitored" (Müller & Guido 2017, 2). In *unsupervised learning* (b), there is no need to prepare the data and the algorithm independently searches for patterns in the input data set. Unsupervised learning is used when no clear target values are specified (Patel 2019, 7–8), e.g., for anomaly detection when analyzing financial transactions to identify fraud attempts, or for recommendation systems that generate suggestions for music, films, or purchases based on the user's preferences, but also for text analysis, such as identifying certain topics or trends in a collection. In the learning style of *reinforcement learning* (c), the AI is programmed to develop a strategy to increase the rewards it receives because of its interaction with an environment. If a strategy is rewarded, the probability of reusing it increases; if the strategy is not rewarded or negatively sanctioned, the probability decreases. It is essentially a trial and error process (Sutton et al. 2018). Reinforcement learning is used, for example, to learn games such as Go or chess, where victories are used as positive reinforcers and defeats as negative ones in model development. However, the evaluation possibilities of output, e.g., in ChatGPT, also flow into the improvement of the models in the same way.

Machine learning is the fundamental basis for the development of *Large Language Models,* AI models based on *Natural Language Processing* (NLP). NLP trains these models to analyze text data and make useful predictions for the next section. Large language models calculate correlations based on their training data and the most likely text output for the respective user input. This process often results in plausible texts that are cohesive on the linguistic text surface, but which – due to the limited *common ground* and world knowledge – can create an "illusion of coherence" (Lotze 2016), which is why they have been given the name "stochastic parrots" (Bender et al. 2021).

However, the probabilities for the output are not calculated at word level, but at token level, as it allows matching grammatical endings or punctuation marks to be predicted. Tokens are meaning bearing units that are produced through the process of tokenization (Michelbacher 2013, 8). In GPT-3.5, tokens comprise an average of four letters, which, according to OpenAI, corresponds to around ¾ of an English word (OpenAI 2023).

As an example for GPT-3, Fig. 1 shows that the token with the highest probability is not always selected. The reason is the *temperature* parameter, which is a setting variable determining the creativity or randomness of the model. The lower the temperature is set, the higher the probability of selecting the most likely token. A temperature set to the lowest value 0 would therefore always respond to a given input with the same, most probable, output. When asked for a suggestion for a name of a discipline that deals with the impact of AI on theology, GPT-4 suggests "Theological AI Studies" at a low temperature, but "Divine Datology" at a high temperature.

The examples from OpenAI's GPT family (the acronym GPT stands for *Generative Pretrained Transformer*) and LaMDA *(Language Model for Dialogue Applications)* use a transformer architecture, a special form of neural network developed by *Google* in 2017, to generate the probabilities (Vaswani et al. 2017; Collins & Ghahramani 2021). This architecture enables a faster and more precise analysis of the input by directing the probabilities for the tokens to be generated to certain semantic spaces by entering the prompt – the text-based input into an AI system – as well as the preceding conversation in the case of dialog models (so-called *attention mechanism*). For example,

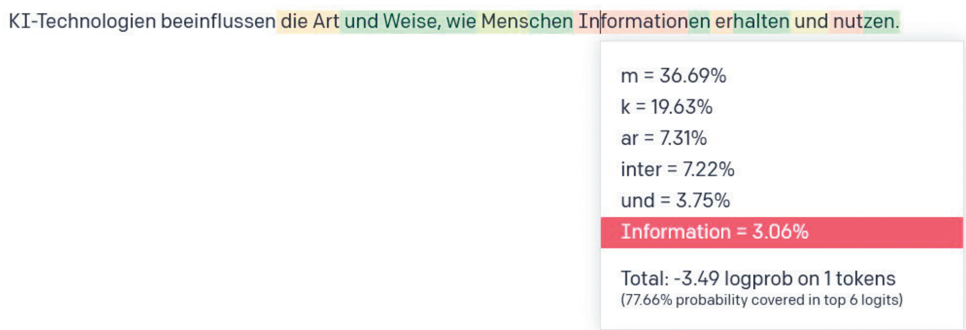


Fig. 1 Probability distribution for the token in the OpenAI Playground. The generated text presented in this figure translates as: "AI technologies are influencing the way people receive and use information."

if the word "tower" is used in an input, then it is decisive for the calculation of the probability for the output whether the input or the preceding discussion refers to chess, castles, or medieval crimes: attention is directed to the respective semantic space in which the most probable output is searched. The field of application of this architecture – developed for the improvement of machine translations, whose shortcoming was the lack of consideration of the context often leading to translations that distorted the meaning – has been considerably expanded and now ensures the higher quality and more accurate output of current systems (Linde 2023).

Even though OpenAI's GPT language models bear the underlying properties of a modern LLM in their name *(**G**enerative **P**retrained **T**ransformer),* other current models are also based on this foundation. *Generative* means that something new is being created. *Pretrained* refers to the training procedure of the LLM, i.e., the underlying machine learning process. The training data for the LLMs comprises a total of around one trillion words (Saboo & Kublik 2023) and, in the case of GPT-3.5, is based on five data sets (Common Crawl, WebText2, Books1, Books2 and Wikipedia) (Brown et al. 2020). LaMDA, on the other hand, is based on English-language dialogs (Collins & Ghahramani 2021), thus giving rise to a central and discussed problem for text-generating AIs, as the training data always contains inherent values and views of the training material. The more frequently material on a topic or concept appears in the training data, the more likely it is that this will be reproduced in the output, which leads Draelos (2023) to conclude: "Large language models (LLMs) like ChatGPT are racist, sexist, homophobic, and in general packed full of all of the worst of society's biases, because they are trained on biased data." The fact that this reality is not so noticeable in the everyday use of ChatGPT is because it is not the raw output of the LLM that is presented, but the output filtered by a *Moderation API* that filters out unwanted output (Markov et al. 2022). However, this detour is susceptible to *prompt injections,* which attempt to persuade systems to make statements that the system should not actually publish (Liu et al. 2023). Draelos therefore is advocating for a focus on the development of LLMs that have a weaker bias. Examples of models that attempt to take this approach include the Claude language model from the US company Anthropic, which purports to have based its model training on the values enshrined in the US Constitution, and the Luminous model created by the Heidelberg start-up Aleph Alpha.

3. Using AI Text Generators at Universities

3.1 Effects

ChatGPT becoming freely accessible sparked different reactions in the field of education – in addition to bans in the USA and a few months later in Italy. There were opposing and swift efforts to create guidelines to give teachers orientation for the constructive use of AI in the context of teaching and examinations. In the meantime, a broad consensus emerged that a blanket ban on AI tools in educational institutions and science is not expedient; rather, that strategies should be pursued to integrate the applications into everyday teaching and research (Buck & Limburg 2023). Examination practices should also be adapted to the technological transformation process. The focus should be less on information retrieval and more on skills and personality development (Budde et al. 2023).

The onset of AI-generated information concerns not only the most traditional form of academic knowledge assurance and skill assessment, namely, the (written) text, and thus these technological developments are at the core of the fundamental academic conception. The focus in the creation and assessment of written work needs to change, since soon – or already at the time of publication of this book – quality texts can be created by generative AI systems. While term papers and theses were previously a medium in which students documented their knowledge acquisition and showed that they could work academically, the aim of these assessments may be more of a means of reflecting on a student's own learning process (Weßels 2022; Weßels et al. 2022; Klein 2023). However, the development of writing skills is a prerequisite for the competent use of text generation tools (Klein 2023). To this end, writing should be more strongly "integrated into subject teaching and used as a thinking tool," students should be more involved in providing mutual feedback and both should be supported by writing tutors. Student conferences are also a proven means of promoting exchange between students (Klein 2023).

Mollick & Mollick (2022, 9) see "new possibilities for interaction" in this sense of a constructive use of AI systems in the writing process and illustrate these constructive uses with three concrete tasks in which generative text tools can help to develop skills in the acquisition of knowledge: a) knowledge transfer by asking the language assistant for an explanation of new concepts to be learned and then checking the output for correctness; b) checking the knowledge gain and stimulating critical thinking by the teacher giving a prompt to generate an essay. The learners should then improve the result by interacting with the application. Documenting the process enables the teacher to understand the student's thought processes and decisions. The third use c) is to encourage the critical evaluation of information and the questioning of one's own assumptions.

Weßels (2022) sees generative AI as having the potential of a "personal learning companion" and predicts that examination discussions, the process of text creation in relation to the final product and, if necessary, the inclusion of tools in the writing process will become more important in assessment. This process entails a change in the role of teachers to designers of the learning space. Digital skills will become increasingly important for students (Friedrich & Tobor 2023). Regarding the concrete writing process, Meyer & Weßels (2023) have published initial results from an AI writing workshop, showing how students see the challenges of these applications and would not adopt the unknown text outputs in a term paper (ibid., 244). At the same time, generative text generators were a helpful tool for overcoming writer's block. Students at the *Hochschule der Medien* see potential both in improving performance, especially by saving time, and in personal development, as they learn how to "use technology" and see this as "preparation for their career and future" (Bihlmeier et al. 2023, slides 4–6).

While many application scenarios are being discussed in academic discourse and tested in small pilot studies, and many experts are calling for their integration into academic teaching, current practice looks different. Empirical surveys from May and June 2023 show most of the students surveyed have never used AI for examination purposes (Humboldt-Universität zu Berlin 2023, 3). The terms most frequently mentioned by the students surveyed in connection with AI in higher education include deception, cheating, uncertainty, plagiarism, and deceit (ibid., 6). Clearly, there is a need for clear rules and instructions as well as training on effective handling to maximize the potential a large portion of respondents see despite many negative associations (ibid., 4–6). Similarly, the need for guidance is also evident on the part of teachers, combined with the desire for guidelines and further training on the use of AI in education (Philologenverband NRW 2023; Surovell 2023).

Initial guidance has been developed since the summer 2022 at the Kiel University of Applied Sciences and the RheinMain University of Applied Sciences in the form of declarations of independence (Limburg et al. 2022, 103; RheinMain University of Applied Sciences 2023). They offer the option of explicitly allowing or excluding the use of text-generating AI tools, either with or without labeling. The adaptation is necessary because the term plagiarism does not apply, even if incorrectly used in connection with AI text tools. Nevertheless, instances of plagiarism are unlikely to be a significant issue in the future, given that LLM-generated texts are inherently distinctive and cannot be readily replicated. AI-generated texts do not constitute plagiarism but are rather classified as ghostwriting. Limburg et al. (2022, 101) see the need for a "revision of the concept of plagiarism."

In order to address the uncertainty regarding the origin of a given text, namely whether it was written by a human or a machine, programs have been developed to distinguish between the two based on linguistic features such as stylistic breaks or "AI-typical" expressions. The leading providers of plagiarism detection software are also working on incorporating a function for recognizing AI-generated texts into their programs (Hipold & Weber 2023). However, OpenAI took its own AI text identifier offline in July 2023 because the recognition rate was too low (Sokolov 2023). Weber-Wulff et al. (2023) concluded the same in their study, in which they tested several programs to determine whether texts can be correctly assigned by machines and humans. The results showed that there was a very high number of false positives and false negatives (ibid., 30). Incorrectness was not due to the lack of quality of the detectors, but rather due to the mathematical elements. Reliable identification would therefore not be possible (Wilkins 2023).

In March 2023, a legal statement commissioned by the Ministry of Culture and Science of the State of North Rhine Westphalia provided clarity on legal issues for the first time – taking a major step towards orientation. It is imperative that AI-generated texts be labeled in a manner that is transparent to third parties, delineating which portions were created through the use of AI tools and to what extent. In formulating labeling strategies, it is essential to consider the relevant legislative framework and the license and terms of use associated with the software in question. In specific instances, labeling can be governed by declarations of independence and, in the context of examination regulations, by the aforementioned legislative and licensing considerations. The adoption of AI-generated text does not inherently contravene the tenets of sound scientific practice as long as the main creative input is furnished by the human operator of the tool. Generative AI can therefore only be an aid (Hoeren 2023, 23). The University of Hohenheim, on the other hand, derived from its statutes on good scientific practice, asserts that the use of ChatGPT is dishonest (Vogelgesang 2023, 8), despite not being explicitly formulated.

Plagiarism can be defined as the appropriation of another person's thoughts or ideas, whether intentionally or unintentionally, without proper attribution. In the context of ghostwriting, it refers to the practice of a person other than the stated author writing the text. Different rules apply to AI-generated texts. According to the German Copyright and Related Rights Act (UrhG), AI is not considered an author. Leading scientific publishers have therefore banned AIs as co-authors (Nature 2023; Science 2023). However, if AI-generated text elements are used in the context of a university assignment, these should be identified. The *Modern Language Association* (MLA) and the *American Psychological Association* (APA), which also publish two of the citation styles that are particularly widespread in the humanities (McAdoo 2023; MLA Style Center 2023), suggest how this identification can be carried out. However, the traditional concept of authorship will have to be altered, given that work with AI text generators will increase, making it increasingly difficult to clearly mark in the texts which parts of the text originate from humans and which from the machine (Kullmann 2023).

The first handouts on the use of AI text generators were published in the first half of 2023 for and by educational institutions in German-speaking countries (Mohr 2023). The general thrust is in line with the legal opinion. A ban on generative AI tools and AI in general in education is not considered sensible; rather, the aim is to integrate them sensibly into teaching and research. The University of Applied Sciences Teachers' Association also calls for the possibility of using AI tools in teaching and refers to the freedom of teaching when structuring their use (Hochschullehrerbund 2023). However, the members are calling for access to such tools for everyone, as they see it as part of the educational mission of universities to prepare students for the world of work. However, doing so requires a basic understanding on the part of both teachers and students. The first tutorials to convey these understandings and present practical application scenarios have already been produced (e.g. KI-Campus 2023).

3.2 Examples of Application

Almost every task and every step of the academic writing process, from topic identification to feedback, can now be completed using AI tools, even with very high quality (Limburg et al. 2023, 4; Gröpler 2023, 367). However, critically reviewing the results is critical. Some examples of these uses can now be examined. The selected tools were chosen as representatives for a range of applications with the same or similar functions to reflect the diversity.

To begin, the chapter shows how an AI text tool can be used for research. For this purpose, ChatGPT (GPT-3.5) was asked to output five important German-language publications on the topic of AI and theology. The following three books and their descriptions were among the results:

- "Gott im Netz: Ein Tagungsband zu Theologie und künstlicher Intelligenz" by Andreas Mertin (Ed.) (Publication year: 2019).
- "Digitale Auferstehung: Von der Zukunft des menschlichen Geistes" by Thomas Metzinger and Svenja Flaßpöhler (2018). This book sheds light on the philosophical and theological aspects of AI and its potential impact on human consciousness and spirituality.
- "Gott 4.0: Warum die Zukunft der Religion in der KI liegt" by Thomas D. Williams (2017). In the work, Williams explores the potential role of AI in religious contexts and how it might affect the future of religion.

These three entries stem from two different queries in the same tool. The different formatting already shows that the outputs differ each time. The results also had no overlaps. At first glance, they look reputable, as the author, title and year of publication are all present. If the existence of these publications or the correctness of the information is not checked, then problems arise. For example, there is a monograph with the title “Gott im Netz,” but with a different subtitle. The author and year of publication do not match the chatbot’s edition.

However, there is an author named Andreas Mertin, who also publishes in the field of religion. The author of the second book also exists, but a book by this name cannot be found. The third book also has no real equivalent, even though there is an author of this name and a monograph “Gott Version 4.0.” Although the results may initially prove that the tools are useless for in-depth research, they can be used to practice critical handling of sources or to become aware of previously unknown authors. The output could also provide ideas for further search terms. However, conventional information infrastructures should always be used for trustworthy sources. Nevertheless, in this domain, a considerable number of AI-supported alternatives and research instruments have emerged, facilitating scientific data discovery through semantic searches.[4] This example illustrates the misleading characteristic of LLMs to conjure up fake results and give the appearance of high plausibility through the coherent formalities, which makes the critical handling of AI text generators imperative.

The following example shows the interaction with a PDF using PDFGear[5] (Fig. 2), which can be downloaded for free. The integrated PDFGear Copilot extends the usual functions of a PDF viewing and editing program with a chat assistant based on GPT-3.5 and allows interaction with the text. To quickly find out which positions are represented in the articles, one prompt could be: “Is digitalization viewed critically in connection with theological issues?”

The tool provides a concise answer, the length and structure of which depends on the parameters of the input. One can also output a list of the main statements of the entire book or just a single chapter, structured by indentions. Here, the output must be critically checked and can then be used in teaching. For example, one task could be to have a student summarize a text and then have a summary created automatically and compare the two to see whether similar core theses have been worked out. For students, teachers, and researchers, a PDF evaluation tool can also help speed up the initial review of electronic literature or to open challenging passages. This function is also integrated into many *AI research assistants* like *Jenni.ai.* As with paraphrasing tools such as *Quillbot,* there is a risk of plagiarism concealment with this application if texts are reformulated by the AI and therefore not recognizable as such. However, a discussion or reflection on the texts can prevent this.

4 Cf. a structured and regularly updated selection at https://www.vkkiwa.de/ki-ressourcen (Accessed: 18 June 2024).

5 PDFGear is representative of other tools that have similar functions, such as ChatGPT Pro with plug-in, BingChat, or ChatPDF.

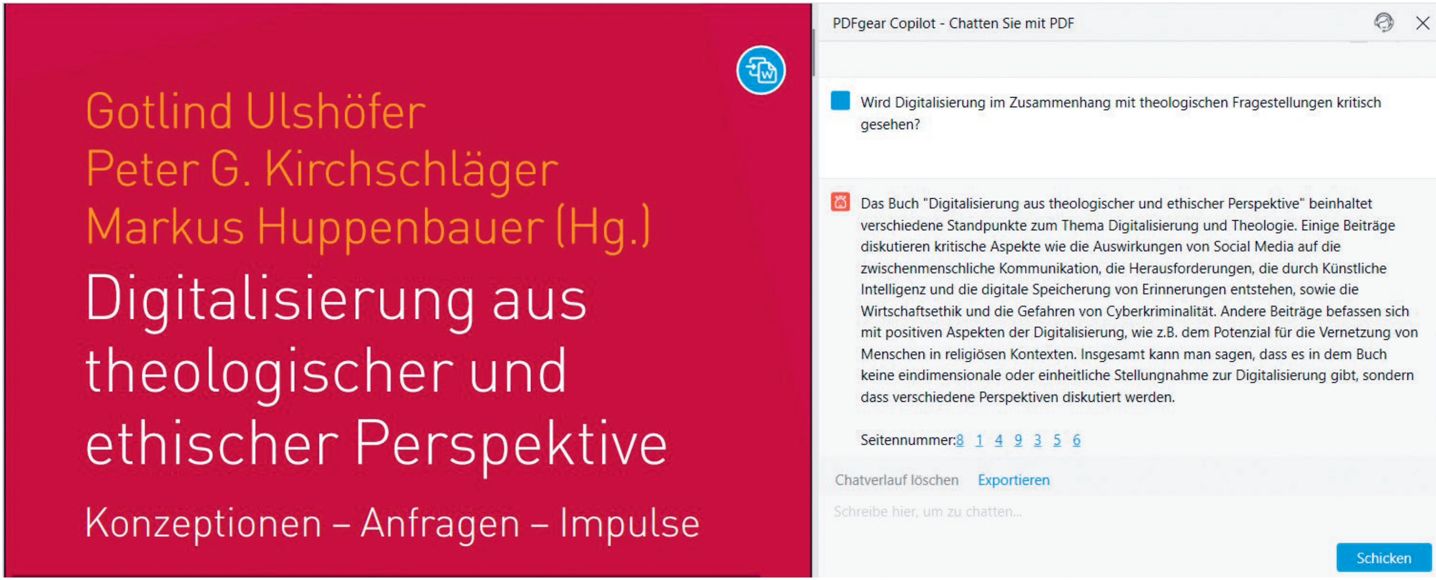


Fig. 2 Chatting with a PDF document*

* For a translation of the chat history shown in Fig. 2 see here:
Prompt: Is digitalization viewed critically in connection with theological issues?
AI-generated answer: The book "Digitalisierung aus theologischer und ethischer Perspektive" ["Digitalization from a theological and ethical perspective"] contains various viewpoints on the topic of digitalization and theology. Some contributions discuss critical aspects such as the impact of social media on interpersonal communication, the challenges posed by artificial intelligence and the digital storage of memories, as well as business ethics and the dangers of cybercrime. Other contributions deal with positive aspects of digitalization, such as the potential for connecting people in religious contexts. Overall, it can be said that there is no one-dimensional or uniform position on digitalization in the book, but that different perspectives are discussed.

Einige Fragestellungen, die zu dem Thema Generative KI und theologische Forschung zusammenhängen könnten sind:

- Wie kann man den Begriff "Generative KI" im Kontext der theologischen Wissenschaft verstehen?
- Welche Möglichkeiten bietet das Verwenden von Generativer KI für die Erforschung des menschlichen Geistes?
- Inwiefern kann die Generative KI dazu beitragen, unser Wissen über Gott zu erweitern oder zu vertiefen?
- Gibt es ethische Bedenken bezüglich der Anwendung von Generativer KI auf dem Gebiet der Theologie?
- Wie können wir sicherstellen, dass die durch Generativer KI erzeugten Ergebnisse mit den Prinzipien und Lehren der Theologie übereinstimmen?

Es ist wichtig zu beachten, dass diese Fragestellungen lediglich als Vorschlag dienen und nicht zwangsläufig alle relevanten Aspekte abdecken. Es ist empfehlenswert, eine breite Palette an Fragen zu erstellen, um ein möglichst detailliertes Verständnis des Themas zu gewinnen.

Fig. 3 Personal screenshot of the chat response from *Open Assistant**

* For a translation of Fig. 3 see here:
Some questions that could be related to the topic of generative AI and theological research are:

- How can we understand the term "generative AI" in the context of theological science?
- What possibilities does the use of Generative AI offer for the exploration of the human mind?
- To what extent can generative AI contribute to expanding or deepening our knowledge of God?
- Are there ethical concerns regarding the application of generative AI in the field of theology?
- How can we ensure that the results generated by Generative AI are in line with the principles and teachings of theology?

It is important to note that these questions are merely suggestions and do not necessarily cover all relevant aspects. It is advisable to create a broad range of questions in order to gain as detailed an understanding of the topic as possible.

As mentioned, text generators can help a writer brainstorm when one finds it difficult to start writing or find a topic. On request, *Open Assistant* creates a selection of possible research questions on the topic of "generative AI in theological research" (Fig. 3).

In the end, one should not simply rely on the result. The results are not always solid output. If the same prompt is entered repeatedly, the answers are very different and do not always correspond to the instructions, which are short and consist of three questions. One can create a research plan based on one of the provided questions or one's own question, even one adapted from actual circumstances.

At a basic level, AI chatbots are also suitable for scrutinizing one's own drafts. One example of this editorial oversight can be seen in the comments that the Claude-2 model produced in response to the request to provide feedback on the section on "How AI text generators work" from this chapter (Fig. 4).[6]

6 Since May 2024, the Claude-3 model family has been regularly available in Europe in three variants: Haiku, Sonnet, and Opus. Additionally, since June 2024, the Claude 3.5 Sonnet model and the

Claude-2-100k Poe

Hier ist mein kritisches Feedback zu deinem Text:

- Der Text ist insgesamt sehr ausführlich und detailliert. Allerdings könnte er stellenweise etwas kürzer gefasst und prägnanter formuliert werden. Einige Abschnitte enthalten redundante Informationen oder gehen zu sehr ins Detail.
- Die Gliederung ist gut nachvollziehbar, könnte aber noch übersichtlicher gestaltet werden. Zwischenüberschriften für die einzelnen Abschnitte würden helfen, den Text zu strukturieren.
- Die Verwendung von Fachbegriffen ist an einigen Stellen sehr technisch. Für den Laien wären einfache Erklärungen und Definitionen der Begriffe hilfreich.
- Inhaltlich werden die wesentlichen Aspekte von KI und maschinellem Lernen abgedeckt. Allerdings könnte stärker auf konkrete Anwendungen und Beispiele eingegangen werden, um die eher theoretischen Ausführungen zu veranschaulichen.
- Die Quellenangaben im Text sollten einheitlich als Fußnoten oder in Klammern im Text erfolgen. Die vollständigen Literaturangaben fehlen noch.
- Insgesamt liest sich der Text sehr sachlich und objektiv. Eine kritischere Diskussion der Chancen und Risiken von KI wäre jedoch wünschenswert. Hier könnten kontroverse Standpunkte stärker herausgearbeitet werden.
- Der Schreibstil ist an manchen Stellen etwas umgangssprachlich. Eine durchgehend sachliche und wissenschaftliche Ausdrucksweise wäre empfehlenswert.

Ich hoffe, diese Anmerkungen sind hilfreich, um den Text weiter zu verbessern. Er bietet eine gute Grundlage mit vielen interessanten Informationen zu KI. Mit einer Straffung, besseren Strukturierung und kritischeren Diskussion ließe sich daraus eine sehr gelungene Abhandlung machen.

Fig. 4 Text criticism with the help of an AI-supported chat tool*

* For a translation of Fig. 4 see here:
Here is my critical feedback on your text:

- Overall, the text is very extensive and detailed. However, it could be shortened in places and formulated more concisely. Some sections contain redundant information or go into too much detail.
- The structure is easy to follow, but could be made even clearer. Subheadings for the individual sections would help to structure the text.
- The use of specialist terms is very technical in some places. Simple explanations and definitions of the terms would be helpful for the layperson.
- In terms of content, the main aspects of AI and machine learning are covered. However, more attention could be paid to concrete applications and examples to illustrate the more theoretical explanations.
- The references in the text should be uniformly given as footnotes or in brackets in the text. The complete references are still missing.
- Overall, the text reads very soberly and objectively. However, a more critical discussion of the opportunities and risks of AI would be desirable. Controversial viewpoints could be highlighted more strongly here.
- The writing style is somewhat colloquial in some places. A more factual and scientific style would be advisable.

I hope the comments are helpful in further improving the text. It provides a good basis with a lot of interesting information on AI. With streamlining, better structuring and more critical discussion, it could be turned into a very successful treatise.

The feedback relates to both content and form, although the comments related to content remain superficial. Nevertheless, AI offers suggestions for further work on the text. One must ensure that the prompt clearly informs the model of the desire for “critical feedback” or that weak points or contradictions should be indicated to obtain the most concrete results. As explained, precise prompts determine where attention should be focused when generating the output and are therefore the central key to useful output. There are now several prompt guides for use in an academic context (e.g., Lenk-Ostendorf 2023). For a final linguistic correction of texts and presentations, specialized applications such as *DeepLWrite*[7] or one of the chat tools based on an LLM can be used.

4. Challenges for Universities, Research and Teaching

The use of generative AI tools is presents opportunities as well as risks. These tools have the potential for the (partial) automation of standard tasks at universities, in teaching and research, such as the preparation of research funding applications or literature reviews, so that more time and focus can be placed on creative work.

However, there are still many challenges. In addition to bias, which, as described above, cannot be completely ruled out by the training data, can be minimized by the most careful curation possible, as with open-source programs (e.g., *Open Assistant* or open source models available on the US developer platform). There are pressing legal issues in the areas of copyright, personal rights, and data protection. The most known and powerful applications on the market not only require the input of personal data, at least an e-mail address, but also use the prompts to train the language model. One can now, at least in the free version of ChatGPT in the free version, exclude this requirement in the settings. Initial approaches to create a protected space for students are already in place. For example, German companies that use the interfaces to the proprietary American language models provide their own user interfaces, which blurs the use of one’s personal data. At *Fobizz*, a teacher can set up a classroom in which learners can test out AI tools (Fobizz 2023). With HAWKI, from the University of Applied Sciences and Arts Hildesheim/Holzminden/Göttingen, university members can log in with their university accounts and securely access the interface to OpenAI (HAWK 2023).

Currently, LLMs remain the most reliable, yet are critical for use in sensitive areas due to a lack of transparency regarding training data, poor working conditions

Claude Artifacts feature have been available. The latter can be used to create and execute interactive diagrams or HTML websites, for example, without programming knowledge.

7 See https://www.deepl.com/de/write (Accessed: 18 June 2024).

for the model's human trainers, and data protection. The possibility of this type of use is due to the large investments from large corporations.

A European solution is also being sought with *OpenGPT-X* (Luber 2023), *Aleph Alpha,* and *OpenAssistant,* among others, to remain competitive and not become further dependent on the US and Chinese AI models currently dominating the market. *OpenAssistant* is a project that aims to counterbalance the big players. Founded in Switzerland, both source code and training data are available on *Github* (Köpf et al. 2023). With LEAM *(Large European AI Models),* a project was launched under the umbrella of the *Bundesverband der Unternehmen der Künstlichen Intelligenz in Deutschland* e.V. (KI Bundesverband) to promote the development and expansion of models that consider “European values and […] standards and regulations” in cooperation with European providers of LLM and representatives from business and science (Bienert et al. 2023, 4;12).

The transformation processes triggered by technological developments in the field of AI will continue to have a considerable influence on the practices of the university system in the future, in teaching, learning and research with the central element of text production, to an extent that cannot even be guessed at yet, and may force us to question our fundamental self-image, at least in some aspects: What does authorship mean? What does good scientific practice mean in the context of the use of AI? What is science and what role do humans play in it? After the initial shocking disruption, science is answering the first sub-questions, gradually providing orientation in the chaos. Many of these questions are likely to become the subject of ongoing controversial discourse. However, the task of research and science is not only to find answers, but above all to ask the right questions. Now, the argument comes full circle, as no one can predict further developments, even in the short term, as the last few months have shown. Between the opportunities and risks roughly outlined as examples, there is a wide range of possibilities for the future of academic study: What is now possible with technology? What do we want to use, how, and for what purpose? What kind of future do we aspire to live, teach, learn, and research in? This range of possibilities now needs to be scientifically explored and positions negotiated in a wider discourse within society.

Bibliography

Adiwardana, D., & Luong, T. (2020). Towards a Conversational Agent that Can Chat About… Anything. In *Google Research Blog.* URL: https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html (Accessed: 18 June 2024).

Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. In *ACM Digital Library, Proceedings of the 2021 ACM*

Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery. DOI: https://doi.org/10.1145/3442188.3445922 (Accessed: 18 June 2024).

Beta Writer. (2019). *Lithium-Ion batteries. A machine-generated summary of current research.* Cham: Springer. DOI: https://doi.org/10.1007/978-3-030-16800-1 (Accessed: 18 June 2024).

Bienert, J., Broch, R., Bunk, P., Burchard, A., Cann, V., Deetjen, B., Dickmann, A., Flores-Herr, N., Friedmann, L., Götz, C., Haenel, F., Hillemann, D., Lambers, K., Lux, L., Otterbach, J., Paass, G., Röhl, C., Schmeier, S., Susallek, P., & Uszkoreit, H. (2023). *Große KI-Modelle für Deutschland. Machbarkeitsstudie.* Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS). DOI: https://doi.org/10.24406/PUBLICA-1047 (Accessed: 18 June 2024).

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, Ch., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, Ch., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). *Language Models are Few-Shot Learners.* Online: arXiv. DOI: https://doi.org/10.48550/arXiv.2005.14165 (Accessed: 18 June 2024).

Buck, I., & Limburg, A. (2023). Hochschulbildung vor dem Hintergrund von Natural Language Processing (KI-Schreibtools). Ein Framework für eine zukunftsfähige Lehr- und Prüfungspraxis, *die hochschullehre,* 9(6), 70–84. DOI: https://doi.org/10.3278/HSL2306W (Accessed: 18 June 2024).

Budde, J., Tobor, J., & Beyermann, J. (2023). Blickpunkt Digitale Prüfungen. *Hochschulforum Digitalisierung,* 1–33. URL: https://hochschulforumdigitalisierung.de/sites/default/files/dateien/HFD_Blickpunkt_Digitale_Pruefungen.pdf (Accessed: 18 June 2024).

Collins, E., & Ghahramani, Z. (2021). LaMDA. Our breakthrough conversation technology. In *Google-Blog.* URL: https://blog.google/technology/ai/lamda (Accessed: 18 June 2024).

Döbel, I., Leis, M., Vogelsang, M.M., Neustroev, D., Petzka, H., Riemer, A., Rüping, S., Voss, A., Wegele, M., & Welz, J. (2018). *Maschinelles Lernen. Eine Analyse zu Kompetenzen, Forschung und Anwendung.* Sankt Augustin/Leipzig/München: Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. URL: https://www.bigdata-ai.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/Fraunhofer_Studie_ML_201809.pdf (Accessed: 18 June 2024).

Draelos, R. (2023). Debiasing ChatGPT, *Medium,* no pag. URL: https://rachel-draelos.medium.com/from-chatgpt-to-puregpt-7124463a100d (Accessed: 18 June 2024).

Friedrich, J.-D., & Tobor, J. (2023). Zur Bedeutung von ChatGPT & der Notwendigkeit eines progressiven Umgangs mit neuen KI-Technologien im Hochschulbereich. Ein Zwischenstand in 6 Thesen. In *Blog of the Hochschulforum Digitalisierung.*

URL: https://hochschulforumdigitalisierung.de/de/blog/chat-gpt-6-thesen (Accessed: 18 June 2024).

Gröpler, J. (2023). Schreiben oder schreiben lassen?, *Forum Bibliothek und Information,* 75(7), 366–368.

Hoeren, T. (2023). Rechtsgutachten zum Umgang mit KI-Software im Hochschulkontext. In P. Salden & J. Leschke (Eds.), *Didaktische und rechtliche Perspektiven auf KI-gestütztes Schreiben in der Hochschulbildung* (pp. 22–41). DOI: https://doi.org/10.13154/294-9734 (Accessed: 18 June 2024).

Klein, A. (2023). "Die Hausarbeit ist tot, es lebe die Hausarbeit!" Entwicklungsorientierung, wissenschaftliches Arbeiten und KI gemeinsam denken. In *Blog of the Hochschulforum Digitalisierung.* URL: https://hochschulforumdigitalisierung.de/de/blog/andrea-klein-KI-Hausarbeit (Accessed: 18 June 2024).

Köpf, A., Kilcher, Y., Rütte, D. v., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., & Mattick, A. (2023). *OpenAssistant Conversations. Democratizing Large Language Model Alignment.* Online: arXiv. URL: https://arxiv.org/pdf/2304.07327 (Accessed: 18 June 2024).

Kullmann, S. (2023). Autorschaft im Schlaglicht. KI und die Regeln der guten wissenschaftlichen Praxis. *Deutsche Gesellschaft für Information & Wissen e. V.* URL: https://dgi-info.de/autorschaft-im-schlaglicht-ki-und-die-regeln-der-guten-wissenschaftlichen-praxis (Accessed: 18 June 2024).

Lenk-Ostendorf, B. *ChatGPT-4 Cookbook.* München: ProLehre | Medien und Didaktik. URL: https://www.prolehre.tum.de/fileadmin/woobtq/www/Angebote_Broschueren_Handreichungen/ChatGPT-4_Cookbook.pdf (Accessed: 18 June 2024).

Limburg, A., Bohle-Jurok, U., Buck, I., Grieshammer, E., Gröpler, J., Knorr, D., Lira Lorca, A., Mundorf, M., Schindler, K., & Wilder, N. (2023). *Zehn Thesen zur Zukunft des Schreibens in der Wissenschaft,* Diskussionspapier, Berlin: Hochschulforum Digitalisierung [= *Diskussionspapier,* 23]. URL: https://hochschulforumdigitalisierung.de/sites/default/files/dateien/HFD_DP_23_Zukunft_Schreiben_Wissenschaft.pdf (Accessed: 18 June 2024).

Limburg, A., Salden, P., Mundorf, M., & Weßels, D. (2022). Plagiarismus in Zeiten Künstlicher Intelligenz, *Zeitschrift für Hochschulentwicklung,* 17(3), 91–106. DOI: https://doi.org/10.3217/ZFHE-17-03/06 (Accessed: 18 June 2024).

Liu, Y., Deng, G., Li, Y., Wang, K., Zhang, T., Liu, Y., Wang, H., Zheng, Y., & Liu, Y. (2023). *Prompt Injection attack against LLM-integrated Applications.* Online: arXiv. URL: https://arxiv.org/pdf/2306.05499 (Accessed: 18 June 2024).

Lotze, N. (2016). *Chatbots. Eine linguistische Analyse* [Diss.]. Berlin et al.: Peter Lang GmbH. [= *Sprache – Medien – Innovationen,* 9]. DOI: https://doi.org/10.3726/b10402 (Accessed: 18 June 2024).

McAdoo, T. (2023). How to cite ChatGPT. In *APA Style Blog.* URL: https://apastyle.apa.org/blog/how-to-cite-chatgpt (Accessed: 18 June 2024).

Meyer, E., & Weßels, D. (2023). Natural Language Processing im akademischen Schreibprozess – mehr Motivation durch Inspiration? Positionspapier basierend auf einer Fallstudie an der Fachhochschule Kiel. In T. Schmohl, A. Watanabe & K. Schelling (Eds.), *Künstliche Intelligenz in der Hochschulbildung. Chancen und Grenzen des KI-gestützten Lernens und Lehrens* (pp. 227–251). Bielefeld: transcript Verlag [= *Hochschulbildung. Lehre und Forschung*, 4]. DOI: https://doi.org/10.14361/9783839457696 (Accessed: 18 June 2024).

Michelbacher, L. (2013). *Multi-word tokenization for natural language processing* [Diss.]. Stuttgart: OPUS. Online Publikationen der Universität Stuttgart. DOI: https://doi.org/10.18419/opus-3208 (Accessed: 18 June 2024).

Mohr, G., Reinmann, G., Blüthmann, N., Lübcke, E. & Kreinsen, M. (2023). Übersicht zu ChatGPT im Kontext Hochschullehre, *Hamburger Zentrum für universitäres Lehren und Lernen (HUL),* 1–12. URL: https://www.hul.uni-hamburg.de/selbstlernmaterialien/dokumente/hul-chatgpt-im-kontext-lehre-2023-01-20.pdf (Accessed: 18 June 2024).

Mollick, E. R., & Mollick, L. (2022). New Modes of Learning Enabled by AI Chatbots. Three Methods and Assignments, *SSRN Electronic Journal* [preprint]. DOI: https://doi.org/10.2139/ssrn.4300783 (Accessed: 18 June 2024).

Müller, A. C., & Guido, S. (2017). *Einführung in Machine Learning mit Python. Praxiswissen Data Science* (K. Rother, Trans.). Heidelberg: dpunkt.verlag GmbH.

Patel, A. A. (2019). *Hands-on unsupervised learning using Python. How to build applied machine learning solutions from unlabeled data.* Sebastopol: O'Reilly Media.

Saboo, S., & Kublik, S. (2022). A Beginner's Guide to GPT-3. In *datacamp blog.* URL: https://www.datacamp.com/blog/a-beginners-guide-to-gpt-3 (Accessed: 18 June 2024).

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning. An Introduction. Adaptive Computation and Machine Learning series. 2. ed. Cambridge, Mass./London: The MIT Press [= *Adaptive Computation and Machine Learning series*].

Ulshöfer, G., Kirchschläger, P. G., & Huppenbauer, M. (Eds.). (2021). *Digitalisierung aus theologischer und ethischer Perspektive. Konzeptionen – Anfragen – Impulse.* Baden-Baden: Nomos Verlagsgesellschaft [= *Religion – Wirtschaft – Politik*]. DOI: https://doi.org/10.5771/9783748924012 (Accessed: 18 June 2024).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need.* Online: arXiv. URL: https://arxiv.org/pdf/1706.03762 (Accessed: 18 June 2024).

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023, 21. Juni). *Testing of Detection Tools for AI-Generated Text.* Online: arXiv. URL: https://arxiv.org/pdf/2306.15666 (Accessed: 18 June 2024).

Wilder, N., Weßels, D., Gröpler, J., Klein, A., & Mundorf, M. (2022). Forschungsintegrität und Künstliche Intelligenz mit Fokus auf den wissenschaftlichen Schreibprozess. Traditionelle Werte auf dem Prüfstand für eine neue Ära. In K. Miller,

M. Valeva & J. Prieß-Buchheit (Eds.), *Verlässliche Wissenschaft – Bedingungen Analyse Reflexion* (pp. 203–223). Darmstadt: wbg academic.

Weßels, D., Mundorf, M., & Wilder, N. (2022). ChatGPT ist erst der Anfang. In *Blog of the Hochschulforum Digitalisierung.* URL: https://hochschulforumdigitalisierung.de/de/blog/ChatGPT-erst-der-anfang (Accessed: 18 June 2024).

Weßels, D. (2022). ChatGPT. Ein Meilenstein der KI-Entwicklung, *Forschung & Lehre,* no pag. URL: https://www.forschung-und-lehre.de/lehre/chatgpt-ein-meilenstein-der-ki-entwicklung-5271 (Accessed: 18 June 2024).

Further References

Bihlmeier, T., Cicek, T., & Hoch, A. (2023). *Studie zur Nutzung von KI-Tools an der Hochschule der Medien [Presentation Slides].* URL: https://filestore.hdm-stuttgart.de/s/kJlLGELhFaHxhg9#pdfviewer (Accessed: 18 June 2024).

Borchers, D. (2022). Zahlen, bitte! 10^9 Speicher voller Modelle. Vom Turing-Test bis ChatGPT. *heise online.* URL: https://www.heise.de/hintergrund/Zahlen-bitte-10-Speicher-voller-Modelle-vom-Turing-Test-bis-Chat-GPT-7392171.html (Accessed: 18 June 2024).

European Commission. (2021). Vorschlag für eine Verordnung des europäischen Parlamentes und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. (2021/0106 (COD)). URL: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_1&format=PDF (Accessed: 18 June 2024).

Fobizz (Ed.). (2023). Digitalen Unterricht einfach und sicher gestalten. Alle hilfreichen Tools & Ki für die Planung und Durchführung von Unterricht an einem Platz | fobizz Tools [Website]. URL: https://fobizz.com/ki-assistenz-fobizz-tools (Accessed: 18 June 2024).

Hipold, P., & Weber, S. (2023). *Erste Plagiatssoftware launcht KI-Erkennung [News release of the University of Innsbruck].* URL: https://www.uibk.ac.at/de/newsroom/2023/erste-plagiatssoftware-launcht-ki-erkennung (Accessed: 18 June 2024).

Hochschule für angewandte Wissenschaft und Kunst Hildesheim/Holzminden/Göttingen – HAWK (Ed.). (2023). HAWKI [AI-chatbot]. URL: https://ai.hawk.de/login.php (Accessed: 18 June 2024).

Hochschule RheinMain. (2023). *Eigenständigkeitserklärung.* URL: https://www.hs-rm.de/fileadmin/Home/Services/Didaktik_und_Digitale_Lehre/Eigenstaendigkeitserklaerung_HSRM_6_23.pdf (Accessed: 18 June 2024).

Hochschullehrerbund (Ed.). (2023). *Wissenschaftsbasierte Lehre und generative KI. Positionspapier.* URL: https://www.hlb.de/fileadmin/hlb-global/downloads/

Positionen/2023-06-12_hlb-Positionspapier_Wissenschaftsbasierte_Lehre_und_generative_KI-Systeme.pdf (Accessed: 18 June 2024).
Humboldt-Universität zu Berlin – HU (Ed.). (2023). *Kurzbefragung zu KI und Prüfungen.* URL: https://pages.cms.hu-berlin.de/doeringn/dashboard/ErgebnisseKurzbefragungKIPruefungen_230621.pdf (Accessed: 18 June 2024).
KI-Campus. (2023). *Sprachassistenzen als Chance für die Hochschullehre | KI-Campus* [Course commentary]. URL: https://ki-campus.org/courses/sprachassistenzen-hochschule (Accessed: 18 June 2024).
Linde, H. (2023). Künstliche Intelligenz. So funktioniert ChatGPT. *golem.de.* URL: https://www.golem.de/news/kuenstliche-intelligenz-so-funktioniert-chatgpt-2302-171644.html (Accessed: 18 June 2024).
Luber, S. (2023). Was ist OpenGPT-X? *BigData-Insider.* URL: https://www.bigdata-insider.de/was-ist-opengpt-x-a-42a8df567400674b7f94418837a80279 (Accessed: 18 June 2024).
Markov, T., Zhang, C., Agarwal, S., Eloundou, T., Lee, T., Adler, S., Jiang, A., & Weng, L. (2023). *New and improved content moderation tooling* [Announcement from OpenAI]. URL: https://openai.com/blog/new-and-improved-content-moderation-tooling (Accessed: 18 June 2024).
MLA Style Center. (2023). *How do I cite generative AI in MLA style?* URL: https://style.mla.org/citing-generative-ai (Accessed: 18 June 2024).
OpenAI. (2023). *Tokenizer.* URL: https://platform.openai.com/tokenizer (Accessed: 18 June 2024).
Philologenverband NRW (Ed.). (2023). *Lehrer-Umfrage zu ChatGPT. Eher Skepsis und Unsicherheit.* URL: https://phv-nrw.de/2023/04/05/lehrer-umfrage-zu-chatgpt-eher-skepsis-und-unsicherheit (Accessed: 18 June 2024).
Science (Ed.). (2023). *Science Journals. Editorial Policies.* Url: https://www.science.org/content/page/science-journals-editorial-policies?adobe_mc=MCMID%3D42627578122959750013493612926978492722%7CMCORGID%3D242B6472541199F70A4C98A6%2540AdobeOrg%7CTS%3D1691159684#authorship (Accessed: 18 June 2024).
Sokolov, D. A. (2023). OpenAI zieht KI-Detektor zurück. *heise online.* URL: https://www.heise.de/news/OpenAI-kann-ChatGPT-Texte-nicht-von-menschlichen-unterscheiden-9227827.html (Accessed: 18 June 2024).
Surovell, E. (2023). Faculty Members Still Aren't Sure What to Make of ChatGPT. *The Chronicle of Higher Education.* URL: https://www.chronicle.com/article/faculty-members-still-arent-sure-what-to-make-of-chatgpt (Accessed: 18 June 2024).
Springer Nature (Ed.). (2023). Artificial Intelligence (AI). In *Nature Portfolio.* URL: https://www.nature.com/nature-portfolio/editorial-policies/ai (Accessed: 18 June 2024).
Virtuelles Kompetenzzentrum KI und wissenschaftliches Arbeiten. (2023). *Virtuelles Kompetenzzentrum KI und wissenschaftliches Arbeiten.* Schreiben lehren und

lernen mit KI. Tools und Techniken für Bildung und Wissenschaft [Website]. URL: https://www.vkkiwa.de (Accessed: 18 June 2024).

Vogelgesang, J., Bleher, J., Krupitzer, C., Stein, A., & Jung, R. (2023). Nutzung von ChatGPT in Lehre und Forschung. Eine Einschätzung der AIDAHO-Projektgruppe [Position Paper]. URL: https://aidaho.uni-hohenheim.de/fileadmin/einrichtungen/aidaho/Dokumente/AIDAHO_ChatGPT_Positionspapier_23-02-09.pdf (Accessed: 18 June 2024).

Wilkins, A. (2023). Reliably detecting AI-generated text is mathematically impossible. *New Scientist*. URL: https://www.newscientist.com/article/2366824-reliably-detecting-ai-generated-text-is-mathematically-impossible (Accessed: 18 June 2024).

Figure Credits

Fig. 1: Personal screenshot from https://platform.openai.com/playground?mode=complete (Accessed: 18. June 2024).

Fig. 2: Personal screenshot of the chat with the e-book PDFgear.

Fig. 3: Personal screenshot of the chat response from Open Assistant.

Fig. 4: Personal screenshot of the chat response from Claude, provided by poe.com.

Reviews of Digital Resources

Ulrike Henny-Krahmer

https://orcid.org/0000-0003-2852-065X

Abstract Research findings in the humanities are traditionally discussed, evaluated, and categorized in reviews of text publications. However, due to the digital transformation of society and science, many other digital forms of research results are emerging, such as digital editions, data collections, and tools, which have rarely been reviewed. In the Digital Humanities, there are several initiatives for reviewing digital resources with the aim of including them in the academic discourse in the same way as text publications and thus helping to harmonize academic evaluation systems with the practice of digital work.*

Keywords Reviews, Evaluation, Digital Editions, Digital Text Collections, Tools

Reviews are an established form of discussing, evaluating, and classifying research results, especially in the humanities and social sciences. Traditionally, reviews have several functions. For one, they provide information about new publications and give an overview of the content, making it easier for readers to get started and gain an overview of a topic. Moreover, by providing their own critical opinion on the subject under review, review authors examine the publication's academic quality. In addition, the reviewed text is further contextualized, i. e., situated within the larger subject area and academic discourse. Academic reviews are usually short and appear in specialist journals or in dedicated review journals.[1] The reviewed works are typically printed products, such as academic monographs, anthologies, or articles, as well as printed editions or encyclopedias.

What is the situation regarding the review of digital resources? Digital resources have long been established, as well, to make research results accessible. In addition to digitally published monographs or articles, digital resources include digital scientific volumes and archives, digital corpora, websites on which research results are presented, sets of research data, software developed in a scientific context, databases, and virtual research environments, as well as multimedia forms of publication

* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

1 In Henny (2018), a sample of 3,000 reviews was taken from the International Bibliography of Reviews of Literature in the Humanities and Social Sciences (IBR-Online). The mean values of three to four pages were then determined.

Ulrike Henny-Krahmer: Reviews of Digital Resources, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 429–441. DOI: https://doi.org/10.11588/heibooks.1521.c21963

such as blog posts or podcasts. This chapter focuses on digital resources that are not analogous forms of printed publications simply created through digitization and, conversely, could not be printed without further work.[2] Their value and importance for academic discourse and knowledge acquisition is undeniable, not only in the Digital Humanities, but also in the humanities in general, as well as other disciplines.

In the age of digital transformation, the research process is fundamentally changing: objects to be researched are being digitized or created directly in digital form, academic communication is taking place primarily in digital and networked form, and research methods are evolving. These changes to the research process are accompanied by a fundamental change in the culture of scientific publication and reception.[3] As a result, the evaluation systems are also undergoing a process of change and adaptation.

This chapter first examines the extent to which the forms of scientific assessment, evaluation, and review have already adapted to the different types of digital research results. The focus is deliberately broader in terms of the contexts and forms of evaluation, while not being limited to reviews, as fundamental questions of evaluating digital research affect all forms of evaluation. The chapter then presents existing initiatives for reviewing digital resources in the Digital Humanities. Finally, the article concludes with an assessment of the importance of reviews of digital research resources and an outlook on their potential for further development.

1. On the Status of Evaluating Digital Research Results

Reflection and debate about the evaluation of digital scholarship began with the onset of the Digital Humanities. Nyhan (2020, 167–176) traces these movements back to the 1960s, when the field was still known as Humanities Computing. However, the idea did not gain traction until the early 2000s, when a series of recommendations were published on how digital research can and should be evaluated, including the “Guidelines for Evaluating Work in Digital Humanities and Digital Media” (*Modern Language Association* 2012), which were first adopted by the MLA in 2000. Since then, numerous handouts and catalogs of criteria for evaluating the results of digital scholarship have been published, ranging from general recommendations for the Digital Humanities as a whole to criteria for certain types of resources (such as scholarly digital editions) and suggestions for specific subject areas and disciplines (most re-

2 One promising area of research is whether more data-oriented scholarly output has been reviewed less or differently than narrative or reporting text forms, i.e., to what extent critical text editions, dictionaries, encyclopedias, catalogs, illustrated volumes, index volumes, or atlases have been reviewed differently than research articles or scholarly monographs, even in print.

3 Jannidis (2023, 1–13) describes these processes of change in digital literature studies.

cently, the “Handreichung zur Rezension von Forschungssoftware in der Archäologie und den Altertumswissenschaften” by Homburg et al. 2020).[4] In addition to recommendations, criteria, and guidelines, there is also research literature explicitly devoted to the topic of evaluating digital research and its results. Two volumes from the 2010s should be mentioned: a section of the journal *Profession* entitled *Evaluating Digital Scholarship* (Schreibman, et al. 2011) and an entire issue of the *Journal of Digital Humanities,* which is introduced by the editors under the title *Closing the Evaluation Gap* (Cohen & Troyano 2012). The series of recommendations and research on the topic of evaluating digital research expresses the need for close attention to the topic. There have been serious points for discussion and clarification, not to mention a concern to provide practical assistance through handouts and to make recommendations for researchers, reviewers, and decision-makers. Current publications and academic events indicate the need for further research on the subject.[5]

Research results in the digital age are undoubtedly no longer only published in the form of traditional text and image publications. Publishing research data and research software is strongly recommended and demanded by funding institutions.[6] The overall awareness of the importance of the sustainable publication of datasets has increased since the formulation of the FAIR principles (“findable, accessible, interoperable, reusable,” Wilkinson et al. (2016), see also the chapter by J. Apel in this volume, pp. 395–396), which are aimed at increasing the reusability and sustainability of research data, especially for machine processing. Nevertheless, what are the challenges?

The report of the European Association of Academies on the sustainable sharing of data in the humanities emphasizes, for example, that an effort by all involved in the research sector is needed to change practices in the handling of research data. A systemic change is needed, a paradigm shift, and a cultural change:

> Universities, research centres, academies, policy makers and funding bodies must review their evaluation methods in order to promote adhesion and commitment to the principles and practices that underpin FAIR data management, because, particularly at these early stages, researchers, data stewards, IT professionals, librarians and archivists, and many others in the research ecosystem need certainty that their involvement will be perceived

4 A list of recommendations for evaluating digital research results can be found in Henny (2018).

5 See, e.g., the article “Nach den Büchern: Rezensionen digitaler Forschungsressourcen” (Neuber & Sahle, 2022), or the session “Attribution and Assessment of Digital Humanities Outputs,” which took place on 16 November 2022 as part of the virtual symposium “Building Digital Humanities” and was organized by the *Digital Humanities Research Initiative.*

6 See, e.g., the report “Sustainable and FAIR Data Sharing in the Humanities” by the *European Association of Academies ALLEA* (2020), or the “Handreichung zum Umgang mit Forschungssoftware” (Katerbow & Feulner 2018), which is supported by the *Digital Information Priority Initiative of the Alliance of German Science Organizations.*

> and recognized in ways that are beneficial to assessment and career progression (ALLEA 2020, 37).

There must be an incentive to disseminate research findings not only in the form of a published text or print publication, but also in the form of digital resources. One significant incentive for doing so is the formal recognition of the work and achievement associated with it. The fact that this recognition for digital research results is not given in the traditional academic system with its established publication forms and channels is one of the core topics of the discussion surrounding the evaluation of digital research results. Efforts to change this lack of recognition are currently being undertaken by some scientific organizations. Last year, on the initiative of the *European University Association* (EUA), *Science Europe,* and the European Commission, an agreement was reached to reform the evaluation of research, which has already been signed by over 600 organizations worldwide at the end of July 2023. The agreement highlights, among other things, the range of types of scientific output and emphasizes that these should be considered when evaluating scientific work:

> Recognise the diversity of research activities and practices, with a diversity of outputs, and reward early sharing and open collaboration […]. Consider also the full range of research outputs, such as scientific publications, data, software, models, methods, theories, algorithms, protocols, workflows, exhibitions, strategies, policy contributions, etc. (Coalition for Advancing Research Assessment 2022, 4).

The addressees of the agreement are those who assess research to make funding decisions, award prizes, define future strategies, and research directions and recruit researchers. The agreement emphasizes the importance of qualitative evaluation and peer review, i. e., the assessment of research by other researchers.

It is therefore clearly stated both by humanities scholars working digitally and by academic organizations and funding institutions that digital research results should be given greater consideration in the evaluation of academic work.

With regard to textual studies and the development and use of virtual research environments, Tóth-Czifra writes of “pressing reharmonization efforts of research evaluation and novel research practices” (Tóth-Czifra 2021): new research practices, in which digital tools are fundamental and produce digital research results, have found their way into the work of academics, but the evaluation criteria and systems have not yet been sufficiently adapted to the new practices.

The reason for this lack of adaptation is, among other things, the nature of the objects to be evaluated and how they are published. Digital objects have different characteristics than printed texts, so that proven evaluation procedures cannot be transferred directly. An important question is the timing of the evaluation of a digital research result. In the case of a traditional print publication, there may be a review before publication and a review after publication. In the digital world, publications can be easily updated. Errors can be corrected on an ongoing basis and deletions or additions can be made at any time, which implies that digital research results can be published earlier and then continuously in different versions. While still common practice to publish a finalized version of text publications, data sets, research software, or complex objects such as digital scholarly editions are often published as alpha or beta versions, in order to be completed on an ongoing basis. Evaluation can therefore take place at any time, considering the variability of the evaluated objects. While also possible for successive editions of printed products, if they contain updates, the dynamics are quite different. New versions of printed publications do not usually appear with the same frequency as new versions of digital publications.

A further issue is the question of which aspects of the digital research results should be evaluated. A digital corpus, an analysis tool, or a digital edition each have a content level that can be evaluated, but also a methodological and a technical level. These levels are often not created by individuals, but by teams in which each person contributes certain expertise. This results in just as many valuation perspectives. An evaluation can also be carried out from a developer or user perspective. Methodological and technical aspects are essential for digital research results, as they make a decisive contribution to their quality and sustainability. Questions concern, for example, the nature of data structures and source code, the use of technical standards, input and output formats, search options and user paths, and all of which go beyond the requirements from the humanities. Thus, new demands are placed on evaluators and reviewers when various aspects of digital resources are to be evaluated. Recommendations and handouts for the evaluation of digital research provide support.

In practice, however, digital resources are still rarely reviewed. Neuber & Sahle (2022) indicate, for example, that although there are review sites for the historical sciences in which digital resources are also discussed, it is done to a much lesser extent than reviews of book publications:

> To date (May 2022), 176 reviews of “websites and databases” and 163 reviews of “digital media” (e.g., CD-ROMs) have been published on H-Soz-Kult. These totals compare to more than 18,000 book reviews […]. Traditional printed journals contain very few reviews of digital research achievements and results. The results are not different even in digital journals (Neuber & Sahle 2022).

Even in Digital Humanities journals, there are hardly any reviews of digital resources. The German-language *Zeitschrift für digitale Geisteswissenschaften* (ZfdG) publishes articles and *working papers*, however, reviews have not yet appeared. The journal *Digital Humanities Quarterly* (DHQ) most recently published on the topics of *Critical Code Studies* and *Tools Criticism* (vol. 17(2), 2023); however, the contributions are academic articles and not reviews in the narrower sense. DHQ also publishes *Reviews:*

since 2007, 47 book reviews, four reviews of conferences and workshops, and three reviews of digital resources have been published. The journal *Digital Scholarship in the Humanities* (DSH) has published 252 reviews since 1986, including 235 book reviews and 17 review articles, according to *Advanced Search,* wherein 13 of the review articles also contain book reviews, a conference report, one review of an exhibition, and one review of a digital commentary, as well as one that is not a review article. Even in relevant Digital Humanities journals, book reviews are more common than reviews of research data, tools, and other digital research results, which shows that established academic publication, evaluation systems, and cultures take time to change. Nevertheless, or perhaps precisely for this reason, there are several initiatives for reviewing digital resources in the Digital Humanities that stand outside the traditional journals and some of which are discussed here in the following.

2. Initiatives for Reviewing Digital Resources in the Digital Humanities

2.1 RIDE – A review journal for digital editions and resources

The review journal RIDE[7] has been publishing review articles on scholarly digital editions, digital text collections, corpora, and on tools and environments for digital scholarly editing since 2014. As of today, 88 review articles have been published in 17 volumes, mostly on digital scholarly editions, which form the focus of the journal. The journal is published by the *Institut für Dokumentologie und Editorik* (IDE) and is an exclusively digital publication in which the articles are published directly in open access in HTML and PDF format.

The starting point for the reviews in the journal are catalogs of criteria written by the members of the IDE and the RIDE Editorial Board, which are intended to help "negotiate digital resources in academic discourse, establish best practice, and advance the discussion of methods" (Sahle et al. 2014). To date, three criteria catalogs have been published for the journal's three sections: the "Criteria Catalog for Reviewing Digital Editions" (Sahle et al. 2014), the "Criteria for Reviewing Digital Text Collections" (Henny & Neuber 2017), and the "Criteria for Reviewing Tools and Environments for Digital Scholarly Editing" (Sichani & Spadini 2018). The catalogs detail the requirements for evaluating digital resources and which aspects should

7 Cf. https://ride.i-d-e.de (Accessed: 19 June 2024). The author of this chapter has been involved in the publication of RIDE from the outset and has been one of the journal's managing editors since 2019. As there is a direct insight into the procedures for reviewing and editing the reviews, RIDE is presented in more detail than the other DH initiatives for reviews of digital resources, which can only be presented from an external perspective.

be centrally discussed. In addition to content-related, edition-philological, and corpus-related criteria, the areas of data modeling, technical implementation, and presentation on the web or user interface are also addressed. Doing so makes it clear that these reviews are a fundamental part of scholarly work on digital editions, text collections, and tools. At the same time, reviewers receive assistance in discussing the digital resources from an interdisciplinary perspective, whereby one does not assume that all elements are treated in a review, but only those relevant.

In RIDE, the reviews appear as articles that are accompanied by a "factsheet," in which essential information on the reviewed resource is summarized in tabular form. The source for the factsheets is questionnaires that are additionally filled out by the reviewers when submitting the review. The data collected in this way also makes it possible to carry out comparative analyses for all reviewed items in the respective field. Statistical graphs are available on the RIDE website, and all data and the review texts themselves are also available in a GitHub repository under a Creative Commons license.[8] The data-based view of the reviewed items ultimately goes back to the criteria catalogs, as the questionnaires were developed based on the catalogs. The reviews are changed by the criteria catalogs in comparison to classic reviews since comparable texts and data are highlighted. Nevertheless, the review articles represent individual perspectives on the reviewed editions, text collections, and tools, and are also considerably longer than classic book reviews: most articles comprise around 15 pages and are therefore closer in length to specialist research articles. This length difference is presumably due to the criteria catalogs, which show in detail what can be discussed in relation to a digital resource, while, on the other hand, the greater length is also a direct result of the diversity of the digital objects that the criteria catalogs depict.

The question of how to deal with the variability and temporality of digital resources when they are reviewed is solved in RIDE such that all links cited in the reviews that are not persistent addresses (such as DOIs or PURLs) are archived on the internet.[9] Archiving ensures that the review articles remain traceable even if link targets change or are no longer available. In addition, reviewers are advised to integrate screenshots as illustrations in the review texts if visual aspects are discussed. Changing resources in RIDE have already led to the same item being reviewed several times, creating an "update" of the review. There is also an effort on the part of the editors not to allow the period until the publication of the peer-reviewed reviews to become too long, in that since 2021 reviews no longer appear only in fully completed volumes, but also individually in a *rolling release* procedure.

The reviews in RIDE include digital resources as research objects in academic discourse by discussing them in text publications that appear in a review journal. This method enhances the review as a text type, given that the articles are more extensive than is usual for reviews and as they undergo a review process before publi-

8 Cf. https://github.com/i-d-e/ride (Accessed: 19 June 2024).

9 Cf. http://web.archive.org (Accessed: 19 June 2024).

cation. Both aspects should incentivize authors to write reviews of digital resources. However, the process also implies more work for writing and publishing. This approach is complemented by data-based access.[10]

2.2 Reviews in Digital Humanities

The journal *Reviews in Digital Humanities* has been peer-reviewing digital research since 2020. The editors describe the publication as a "pilot of a peer reviewed journal and project registry that facilitates scholarly evaluation and dissemination of Digital Humanities work and its outputs."[11] Unlike RIDE, the creators of digital resources submit project descriptions, which are then peer-reviewed by other researchers. Both the project description and the review are published in *Reviews in DH.* Submissions can be descriptions of digital archives, multimedia science, digital exhibitions, visualizations, computer games, and software. The journal also encourages submissions on digital scholarship in the fields of critical ethnic studies, African diaspora, indigenous peoples, Latinx, Asian American, and postcolonial studies. The project descriptions and reviews are published under a CC-BY license and are relatively short.[12] Since 2020, 4 annual volumes with a total of 39 issues have been published, each containing around 4 to 5 contributions. Each article is addressable with its own DOI. The submitted projects can also be searched via a "Project Registry" according to 4 criteria (alphabetically, by epoch, by subject area, and by topic or method).

In *Reviews in DH*, the authors of the project descriptions and reviews, as in RIDE, are given information on what can and should be addressed, including both scholarly and technical aspects. Reference is made to external handouts on the evaluation of digital science. The aim of the journal is to promote critical discourse on digital science and to do so in a way useful for other scientists. The journal aims specifically to address the "evaluation gap" resulting from the increasing number and scope of digital projects for which there is not enough opportunity for review in existing journals. At the same time, the importance of peer review for digital research is emphasized, as it is important for the recognition of scientific achievement in the creation of digital resources. "Reviews in DH" therefore takes a different approach than RIDE, as the reviews are deliberately kept short and include project descriptions. The content focus

10 In addition to RIDE, there is also the initiative "Construction KIT: A review journal for research tools and data services" (CKIT), which was created in the context of the NFDI4Culture consortium of the *National Research Data Infrastructure.* The first review article was published in October 2023 in the journal, which focuses on research software. Cf. https://journals.ub.uni-heidelberg.de/index.php/ckit (Accessed: 19 June 2024).

11 See https://reviewsindh.pubpub.org (Accessed: 19 June 2024).

12 In 10 randomized samples pulled, the reviews were between approx. 400 and 1,100 words in length. The journal itself suggests a length of 500 words for both the project descriptions and the reviews.

is also broader than RIDE. All types of digital projects and resources can be submitted. In this way, a large number of projects can be examined. What both journals have in common is their desire to strengthen the review system for digital resources and offer systematic access to the content in addition to texts.[13]

2.3 DHTech Community Code Reviews

The *DHTech Community Code Reviews* is still a young initiative. The initiative reviews the code of software projects from the Digital Humanities. The initiative is backed by a working group of the *DHTech Special Interest Group* established in 2017 and is active under the umbrella of the international *Alliance of Digital Humanities Organizations* (ADHO).[14] DHTech promotes the development and reuse of software in DH and provides a place of collaboration for all those involved in the development of software in this field of research.

The *Code Review* has been active since 2022 and have reviewed 3 tools.[15] Code reviews are a common technique in software development for improving software. The source code of a piece of software is reviewed by someone, not the author of the software, who is himself or herself a software programmer. In an interactive process, the software reviewers ask questions and make suggestions for improvement. In so doing, the quality of a program's source code can be improved by correcting errors and making the code more readable and easier to maintain. DHTech's code reviews are performed directly on the software development platform GitHub. The working group provides information on how a code review can be prepared, carried out, and the criteria according to which code can be reviewed. Developers submit their software for review. The prerequisite for review is a published code with a license, including documentation and sample data. Also, due to time constraints, an extensive code cannot be reviewed, only code that can be understood and reviewed within one hour.[16] A copy of the code is made for the purpose of the review and the code re-

13 Like *Reviews in DH*, the *DHCommons Journal* combined short project descriptions with reviews of the projects. The journal was published from 2015 to 2016 by the international network *centerNet* but is no longer available online. Since 2017, the European project DARIAH-EU has also been publishing the blog "OpenMethods: Highlighting Digital Humanities Methods and Tools" (https://openmethods.dariah.eu, accessed: 19 June 2024). Short blog posts draw attention to curated content on DH methods and tools since there remains little focus on these issues in the DH literature. These issues are important to discuss critically to highlight the value, opportunities, and challenges of DH. Content selected for the blog undergoes a quality check and is categorized according to the TaDiRAH taxonomy, which can then be used to filter the blog's content. The OpenMethods blog takes a minimalist approach to reviewing DH methods and tools.

14 Cf. https://adho.org/sigs/#DHTech (Accessed: 19 June 2024).

15 Cf. https://dhcodereview.github.io (Accessed: 19 June 2024).

16 Cf. https://github.com/DHCodeReview/DHCodeReview/wiki/Authors:-Preparing-a-Code-Review (Accessed: 16 June 2024).

viewers can then comment on it directly.[17] The "Code Review Guidelines" provided by DHTech give instructions for the dialog between developers and reviewers and checklists for essential properties of the source code to be reviewed.[18]

Regarding the review of digital resources, DHTech's *Community Code Reviews* represent a completely different approach than the review journals presented in this chapter. By reviewing source code, they are directly and exclusively addressing the technical level of DH tools. The software is not evaluated and contextualized in a review text; the reviewers engage directly with the developers, with the aim of providing feedback and improving the quality of the digital resource. Due to digital resources being constantly updated, suggestions for improvement from reviewers are also frequently accepted in other cases and implemented for future versions. These are mediated by the review text yet not as directly as with code reviews. The code reviews are a very important initiative, as the work of research software developers has hardly been recognized in traditional scientific and academic discourse. However, the code reviews also take place outside the usual publication channels in the humanities. In contrast to RIDE, *Code Reviews* are not attempting to make the digital resources the object of classic specialist reviews; instead, a form of review is selected that is directly oriented towards the reviewed object and follows the practices of software development.[19]

3. Prospect: On the Importance and Forms of Reviews of Digital Resources

Reviews of digital resources remain far too rare. Although humanities scholars now work with digital data and tools in almost all areas and produce research results in digital form, reviewing is still mainly focused on reviewing traditional forms of scholarly output such as monographs, edited volumes, articles, and other publications that can be traced back to the book form, regardless of whether they appear in print or digitally. Websites with scientific content, scientifically compiled and edited data collections, and digital objects as well as research software have thus far rarely been reviewed in specialist journals in the humanities, even in the DH themselves.

17 Cf. https://github.com/DHCodeReview/DHCodeReview/wiki/Conducting-a-Code-Review (Accessed: 16 June 2024).

18 Cf. https://github.com/DHCodeReview/DHCodeReview/wiki/Code-Review-Guidelines (Accessed: 16 June 2024).

19 Software reviews are also performed by the *Journal of Open Source Software* (JOSS), in which developers can submit short essays on their research software. Both the software and the essay are then reviewed by experts.

Thus, there have been a number of recent initiatives developed in the DH that are specifically dedicated to the review of digital resources, such as the review journals RIDE and *Reviews in DH* presented here or the *Community Code Reviews* within the DHTech Working Group. Different approaches to reviews of digital resources have been taken. On the one hand, digital resources become the subject of reviews in text form, thereby integrating them into the normal academic discourse. The reviews of digital resources retain the essential functions of classic reviews: an overview of the project, an evaluation of the subject matter, and an overall classification of the work into the wider disciplinary context. Even still, the form of the reviews changes when referring to digital objects; they are juxtaposed with project descriptions, accompanied by indexes and data collections, and vary in scope. There have been attempts to simplify the process for reviewers to discuss complex digital resources by providing them with guidelines so that humanities scholars can also take technical aspects into account.

In the future, reviews of digital resources could either be written collaboratively by several authors with different specialist backgrounds, or there could be several reviews written by individual authors, each focusing on a specific layer of the review of the digital resource in question, such as their content, data models and algorithms, the technical implementation, or aspects of the design of the presentation in the digital world. A review of different layers of digital resources helps to disclose and recognize the scientific work of those involved in their production. Reviews can then also take on completely new forms that move further away from traditional text reviews, such as the example given by the code reviews for the assessment of software. It will be important that new forms of review are also recognized as part of the scientific discourse. Regardless of exactly how reviews of digital resources are implemented and designed, the most important aspect is that they are carried out at all and that they provide the digital transformation of science and the academy with an appropriate and functioning practice of evaluation and assessment. Reviews of digital resources can support the transition to an open, digital scientific culture by creating incentives for scientists to continue producing and publishing digital research results. They are a building block in a process of change that also encompasses other areas, such as citation practices, rules for recording scientific publications in reference systems, or the evaluation of research achievements in general.

References

Coalition for Advancing Research Assessment (CoARA). (Ed.). (2022). *Agreement on Reforming Research Assessment.* URL: https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf (Accessed: 19 June 2024).

Cohen, D.J., & Troyano, J.F. (2012). Closing the Evaluation Gap, *Journal of Digital Humanities,* 1(4), no pag. URL: https://journalofdigitalhumanities.org/1-4/closing-the-evaluation-gap (Accessed: 19 June 2024).

Damerow, J., Sutton Koeser, R., Gao, A., Vogl, M., Zandbank, I., Tharsen, J., Casties, R., Westerling, K., & Carver, J. (2022). Establishing a Code Review Community for DH. In *Digital Humanities 2022. Responding to Asian Diversity* (pp. 633–634). Tokyo: DH2022 Local Organizing Committee. URL: https://dh2022.dhii.asia/dh2022bookofabsts.pdf (Accessed: 19 June 2024).

Harrower, N., Maryl, M., Immenhauser, B., & Biro, T. (Eds.). (2020). *Sustainable and FAIR Data Sharing in the Humanities. Recommendations of the ALLEA Working Group E-Humanities.* Berlin: ALLEA. All European Academies. DOI: https://doi.org/10.7486/DRI.tq582c863 (Accessed: 19 June 2024).

Henny, U. (2018). Reviewing von digitalen Editionen im Kontext der Evaluation digitaler Forschungsergebnisse. In R.S. Kamzelak, & T. Steyer (Eds.), *Digitale Metamorphose. Digital Humanities und Editionswissenschaft* (no pag.). Wolfenbüttel: Herzog August Bibliothek [= *Sonderband der Zeitschrift für digitale Geisteswissenschaften,* 2]. DOI: https://doi.org/10.17175/sb002_006 (Accessed: 19 June 2024).

Ead., & Neuber, F. (2017). *Criteria for Reviewing Digital Text Collections.* Version 1.0. In collaboration with the members of the IDE. Online: Institut für Dokumentologie und Editorik. URL: https://www.i-d-e.de/publikationen/weitereschriften/criteria-text-collections-version-1-0 (Accessed: 19 June 2024).

Homburg, T., Klammt, A., Mara, H., Schmid, C., Schmidt, S.Ch., Thiery, F., & Trognitz, M. (2020). Diskussionsbeitrag. Handreichung zur Rezension von Forschungssoftware in der Archäologie und den Altertumswissenschaften, *Archäologische Informationen,* 43, 357–371. DOI: https://doi.org/10.11588/ai.2020.1.81422 (Accessed: 19 June 2024).

Jannidis, F. (2023). Digitale Literaturwissenschaft. Zur Einführung. In id. (Ed.), *Digitale Literaturwissenschaft. DFG-Sympsion 2017* (pp. 1–16). Stuttgart: J.B. Metzler. DOI: https://doi.org/10.1007/978-3-476-05886-7_1 (Accessed: 19 June 2024).

Katerbow, M., & Feulner, G. (2018). *Handreichung zum Umgang mit Forschungssoftware.* Zenodo: Schwerpunktinitiative Digitale Information der Allianz der deutschen Wissenschaftsorganisationen. DOI: https://doi.org/10.5281/zenodo.1172970 (Accessed: 19 June 2024).

Modern Language Association. (Ed.). (2012). *Guidelines for Evaluating Work in Digital Humanities and Digital Media.* URL: http://www.mla.org/guidelines_evaluation_digital?ot=letterhead (Accessed: 19 June 2024).

Neuber, F., & Sahle, P. (2022). Nach den Büchern. Rezensionen digitaler Forschungsressourcen, *H-Soz-Kult. Forum. Buchrezensionen in den Geschichtswissenschaften,* no pag. URL: www.hsozkult.de/debate/id/fddebate-132457 (Accessed: 19 June 2024).

Nyhan, J. (2020). The Evaluation and Peer Review of Digital Scholarship in the Humanities. Experiences, Discussions, and Histories. In J. Edmond (Ed.), *Digital Technology and the Practices of Humanities Research* (pp. 163–182). Cambridge, UK: Open Book Publishers. DOI: https://doi.org/10.11647/OBP.0192.07 (Accessed: 19 June 2024).

Sahle, P. (2014). Kriterienkatalog für die Besprechung digitaler Editionen. Version 1.1. With the collaboration of G. Vogeler and the members of the IDE. Online: Institut für Dokumentologie und Editorik. URL: https://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1 (Accessed: 19 June 2024).

Schreibman, S., Mandell, L., & Olsen, S. (Eds.) (2011). Evaluating Digital Scholarship. In R. G. Feal (Ed.), *Profession* (pp. 123–267). New York: The Modern Language Association of America. URL: https://www.jstor.org/stable/i40080515 (Accessed: 19 June 2024).

Sichani, A.-M., & Spadini, E. (2018). Criteria for Reviewing Tools and Environments for Digital Scholarly Editing. Version 1.0. In collaboration with the members of the IDE. Online: Institut für Dokumentologie und Editorik. URL: https://www.i-d-e.de/publikationen/weitereschriften/criteria-tools-version-1 (Accessed: 19 June 2024).

Tóth-Czifra, E. (2021). Rethinking text, techné and tenure. Evaluation and peer review challenges around Virtual Research Environments in the Arts and Humanities, *Classics@Journal,* 18, no pag. URL: https://classics-at.chs.harvard.edu/classics18-toth-czifra (Accessed: 19 June 2024).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, Ph. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, Ch. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., C't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, Ph., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, Th., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship, *Scientific Data,* 3(1), 1–9. DOI: https://doi.org/10.1038/sdata.2016.18 (Accessed: 19 June 2024).

Transformation of Conventional Research Environments and Publication Forms

Clifford Anderson

Abstract In this paper, I explore how and to what extent transformations in scholarly research and publication have effected changes in scholarship itself. Taking my theoretical cues from media studies, I use the dual lenses of media displacement and media saturation theory to analyze the alteration in practice that has occurred as scholars shift from analog to digital forms of research and writing. We see that this shift was not itself binary but exists along a continuum. For instance, the digitization of primary and secondary sources promised to unlock new methods of digital research, but the often poor quality of optical character recognition impedes the application of those methods. Word processing promised to speed up scholarly production and, in some senses, succeeded but also managed to occlude the digital texts, making them harder to aggregate and repurpose. Web annotations aimed to fulfill the vision of a distributed web of critical commentary but the scale of the internet makes achieving such dreams hard to pull off. Digital tools for enumerative bibliography have largely automated the formulation of citations, though they have not yet broken with the form itself. Finally, digital publishing still relies, by and large, on interfaces that mimic their analog counterparts. In short, we find that digital tools and methods are not displacing the analog but supplementing them. Is this a sign of an ongoing and incomplete digital revolution or a stable and enduring scholarly synthesis?

Keywords Scientific Culture, Media Studies, Digitization, Librarianship, Text Processing, Digital Humanities Pedagogy

There is widespread agreement that Digital Humanities has transformed conventional research and publication. The emergence of blog posts, digital editions, code notebooks, and data repositories alongside the conference talks, articles, and monographs has enriched the range of scholarly outputs. How has this expanded scope of academic outputs changed the nature of research itself? At its root, this is an empirical question that should be addressed through mixed methods research into the changing habits of religious studies scholars. In this contribution, I prepare the ground for such a study by examining the transformations of scholarly research environments in theology and religious studies through the lens of media studies.[1] To analyze these

1 Editor's note: The author uses the terms *theology* and *religious studies* according to the Anglo-Saxon understanding, where religious studies refer to all research that has religion as its object of study (cf. the contribution by Ch. A. Nunn in this volume, p. 71).

Clifford Anderson: Transformation of Conventional Research Environments and Publication Forms, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 443–462.
DOI: https://doi.org/10.11588/heibooks.1521.c21964

transformations, I draw on contested concepts in media studies, namely media displacement theory and the notion of media saturation (Newell et al. 2008). To what extent have new digital media supplanted our conventional research methods? To what degree has our media ecology become inundated with digital technologies?

What is media displacement theory? The core idea is straightforward. We have limited time at our disposal to consume media. If we begin to engage with new forms of media, we need to find the time somewhere in our day. As Bryant & Fondren (2009, 505) remark, "In displacement theory, the core assertion is that media consumption will displace some other activity or activities, such as exercise or social interaction, or even shift time from one medium to another."

Scholarship differs from media consumption, of course, by its focus on production. The purpose of research and publication is to produce and communicate new knowledge. The scholarly activities that support these activities have changed considerably as humanists engage with new forms of digital media. Analogous questions arise about potential displacements. Do scholars spend more time reading ebooks or perusing PDFs online than pulling monographs from library shelves or consulting offprints of journal articles? Has the recent availability of audiobooks from university presses impacted the time spent reading?

The question of saturation is correlated closely with displacement theory. In media studies, saturation indicates the limits of media consumption. As such, the concept resembles the notion of information overload. On its own, information overload is nothing new. As Blair (2011) has argued in *Too Much to Know: Managing Scholarly Information before the Modern Age*, scholars since antiquity have faced the challenge of having too many sources and have designed scholarly technologies, from indexes to florilegia and commonplace books, to mitigate the flow of information. Media saturation theory sharpens the problem of information overload by underscoring that scholarly communication flows through proliferating forms of media. As scholars, we must consider whether or how to combine traditional channels of research – archival manuscripts, journal articles, monographs – with data flowing through newspapers, podcasts, preprint servers, radio, social media, television shows, webinars, websites, YouTube, and now generative artificial intelligence. "The saturation of media tends to result in increasing fragmentation of information," remarks Wasiak (2008, 113), "as one navigates space and media sources." As media proliferates, we cannot connect the threads, tracing conversations from TikTok to blogs to scholarly articles and then back again. As new channels (*Clubhouse*, anyone?) pop up, we may just tune them out because our media environment is already suffused with competitors for our attention. As we shall see, a strong trend in the Digital Humanities resists the proliferation of new media by reducing all media, as far as possible, to data and using command line utilities (Bash, ZSh, etc.) to manipulate those data.

Adopting a media studies lens also helps us to look more holistically at the digital turn in religious studies and theology. On the one hand, theologians and scholars of religion continue to carry out their research and publication in ways similar to fifty years ago. While computation has wholly transformed the sciences, requiring graduate students and postdocs to develop skills in data analysis and large-scale computing to carry out their research, the same cannot be said of theologians. In *The Place of Computation in the Study of Culture,* Daniel Allington distinguishes, in the line of C.P. Snow, two cultures of the academy: the hermeneutical and the empirical. Among what he terms the "*essentially* hermeneutic," he places the discipline of theology (Allington 2022, 374). "And computation has almost nothing to contribute under such a paradigm," he insists, "unless we mean those computations which go on, unnoticed, in the background, incessantly, so that emails can take the place of the postal service and a word processor can take the place of a typewriter. But almost nobody cares about those (ibid., 373)." From a media studies perspective, we may come to care about these unheralded shifts in research and publication practices because, collectively, they add up to qualitative differences in our scholarly media ecology.

1. The Analog to Digital Shift

When we consider the transformation of conventional research environments, the critical driver that comes to mind is the shift from analog to digital research methods and publication forms. At the center of contemporary scholarly communications stands the networked computer, replacing nearly all previous forms of academic technology. Or so the presumption goes.

The transition from analog to digital has eventuated in both subtle and profound transformations in research and publication patterns. Analog methods have not simply given way to digital successors. Pace the so-called *media displacement* theory, analog and digital methods flourish side by side. From an anecdotal perspective, scholars take notes about the book they are reading, underling or highlighting critical passages in pencil; they also mark up articles in PDF format, storing their digital annotations in tools like *Papers* or *Zotero.*

Robert Hassan, Professor of Media and Communication at the University of Melbourne, argues in *Analog* that we should not consider analog and digital as antipodes. The expanse of the analog world is greater and more connected with our personal and cultural histories; analog habits of research have become second nature, or parts of "our extended mind" (to use the language of Clark & Chalmers 1998). Given the millennia of entanglement of our minds with analog tools of scholarship, it should come as no surprise that these habits did not immediately give way to digital equivalents with the advent of *Unix* time on 1 January 1970. "We need to remind ourselves that to write and read is to interact with a technology," contends Hassan (2023, 132). "At a deeper level we need to remind ourselves also that the technology is analog, in that it corresponds, symbolically, to speaking, to hearing the voice and, with it, the mind's thoughts." Hassan acknowledges that just as Friedrich Nietzsche's purchase of an early typewriter (a Hansen Writing Ball) in 1882 may have pushed him toward his late aphoristic style, so may the digitization of our research environments shape the contours of our research. “At a deeper philosophical level, screen-reading suggests a new relationship with knowledge, which is to say, a new interface in the constituting of the reality of the world (ibid., 139).” As we shall see, the interface question looms large over the new world of digital research and publication.

2. Digitization

If you found yourself teleported into a library from fifty years ago, things at first inspection would look generally the same. The same row on row of books ordered (at least in the United States) on shelves according to the Library of Congress classification system would span the floors. You would find students in the reading rooms and carrels, pouring over books and bound periodicals. The differences would strike you when you sought a book on the shelf. To carry out that task, you would need to consult the catalog, held in drawer after drawer of cards that might span an entire room. In the library of the 1970s, you would also find technology that has since disappeared or radically diminished. In the interlibrary loan department, alongside analog phones and Xerox 4000 photocopy machines, for example, you would find a teletype machine used to communicate loan requests from peer libraries.[2] Off the side of the reading room, you would find metal cabinets full of microfilm and microfiche along with the requisite readers; would you know how to thread the spool and advance the reader to the beginning of the article you hoped to read?[3] As for computers, you might find mainframes here and there in the technical services units of large university libraries. Still, you wouldn’t find them anywhere in the public areas. What you might notice at the reference desk is a so-called “dumb terminal,” that is, a system like the DEC VT52 terminal, which connected to the DIALOG Online Search System or a competitor; using a compact query language, the librarian would search multiple databases, carefully avoiding returning too many results as search results were metered (Schatz 1997).

From a phenomenological perspective, the most significant change you would experience relates to the level of intermediation. In the library of the past, librarians played a central role in assisting patrons with navigating to sources of information. The placement of the reference desk at the center of the library both reflected and reinforced the librarian’s mediating function. The past fifty years have seen tremendous

2 See https://www.facebook.com/pasadenalibrary/posts/whats-that-machine-its-a-teletype-machine-this-is-a-teletype-model-28-which-had-/10157896789598049 (Accessed 26 June 2024).
3 See https://edtechmagazine.com/higher/article/2017/05/microfiche-was-dawn-multimedia-research (Accessed 26 June 2024).

disintermediation of research environments, primarily removing librarians from the day-to-day research process.

The first wave of disintermediation made metadata directly available to the public. Librarians began to migrate cataloging records from print to digital format in the 1960s and 1970s. By the 1980s, OPACs *(Online Public Access Catalogs)* had become standard in academic libraries in North America, at first coexisting with and then functionally displacing card catalogs. Partnerships like the *Research Libraries Group* (RLG) and the *Ohio College Library Center* (OCLC)[4] enabled researchers to discover literature in peer library collections; these days, scholars can use OCLC's *WorldCat* both to search for relevant items and to identify libraries that hold those items. The same applies to periodical sources. Readers can readily find metadata about articles of interest through *Google Scholar*, though they may encounter paywalls when seeking to download the PDFs.

The second wave of disintermediation resulted from the digitization of collections. In *Along Came Google: A History of Library Digitization*, Marcum & Schonfeld (2021) document the initiatives to digitize library collections before Google. From the mid-1990s, librarians began forming coalitions to make library collections available online. However, the scale of these ambitions, disagreements about directions, and common pool resources dilemmas inhibited the success of these projects. Starting in 2004, however, the entry of Google changed the equation drastically; the *Google Books* project revived hopes that all books could be made accessible online. Google's scanning process introduced errors with fingers and hands appearing in images occasionally but proved efficient (James 2010), all too efficient for publishers. A flurry of lawsuits led to the retrenchment of Google's planned universal library. As Marcum & Schonfeld (2021, 188 f.) remark, "Rather than a universal digital library, we have a potpourri of digital collections, with greater or lesser access, as well as libraries that have individually become digital, more or less." While the prospects of realizing a universal digital library have diminished, scholars can still find monographs online between these sources, reducing their dependence on the library as a physical center of information.

From a Digital Humanities perspective, digitizing these volumes provides an incredible start but also falls far short of the goal. To conduct data-driven research, scholars need direct access to the underlying data and metadata. In many cases, the intellectual property restrictions bar access to the data in raw form. When it is possible to download or scrape the data, scholars frequently find, to their dismay, that the optical character recognition software has produced a nonsensical mess. Quoting from a recent document, for example, we see this sentence, which is no worse or better than the ones surrounding it: "After a week' s lness, E5= abeth, eldet. dn. of Mr. H. W. By* 1s Mfay. m e Fhebs wif of A. C. 1e_, esq. d f of h" e Be,. Ta. s forrlyo em of s Rea f Uonl." Good luck trying to carry out textual analysis on documents this garbled.

4 RLG merged with OCLC in 2006.

3. Word Processing

As Allington remarked, the most profound shifts from analog to digital research may simultaneously be regarded as the most trivial. To wit, the shift in writing practices. Fifty years ago, theologians generally wrote their texts by hand, then typed them up (or sent their manuscripts by interoffice mail to department secretaries to be typed) for submission to a publisher. These days, nearly everyone writes using word processing software and emails files (in Microsoft Word format) directly to editors.

In *Track Changes. A Literary History of Word Processing*, Matthew Kirschenbaum examines the effects of the switch to word processing on writers and writing through the lens of media studies. He demonstrates that what feels inevitable now struck writers of the period as alien, exciting, and unsettling in equal parts during the 1980s. As writers experimented with word processing systems, they regularly remarked about how these systems would transform literary production. Some, like Stephen King, experimented with word processors early on in the hope that they would accelerate their already prolific output. Others, like Gore Vidal, bemoaned that "the idea of literature is being erased by the word processor (Kirschenbaum 2016, 43)." Reflecting from a contemporary vantage point, Kirschenbaum pronounces a tempered judgment. "*Pace* Gore Vidal, word processing did not erase literature, not in any sense I can fathom. Neither, of course, did it perfect literature," he writes, commenting, "But like the typewriter before it, word processing changed the face of literary culture and our imagination of literary authorship (ibid., 243)." That is, the switch to the word processor altered our image of the author; today, we picture writers (and, by extension, scholars) huddled over laptops in coffee shops rather than pounding out prose on manual typewriters or composing essays by candlelight with quill pins. "See what big letters I make as I write to you with my own hand!" Paul exclaimed in Galatians 6:11 as he stopped dictating to his amanuensis. These days, our handwriting has become so squiggly due to our focus on typing that students prefer faculty to proffer feedback in their learning management systems than to scrawl comments in red on their papers.

Will the current ubiquity of Microsoft Word persist among scholars in the long term? As anyone who has used Word for a decade or longer knows, the project managers at Microsoft did not have researchers in view when designing their word processor. For years, academics struggled with footnotes that appeared on the wrong page and other annoying glitches. While these challenges to using Word for academic writing have generally been overcome, there remains a significant mismatch between Word and the Digital Humanities. The "What You See Is What You Get" (WYSIWYG) model of word processing, which sought to replicate the polished layout of the page even in the phase of composition, impedes the perspective of text as data. Why is this so? Word obfuscates the structure of documents by focusing on the presentation of the text on the screen (and, hence, on the printed page). There are tools in Word to mark out the structure of documents, like primary and secondary headings, etc.

However, authors typically ignore such features when formatting their documents and use visual clues like bolding, italics, and larger fonts to indicate textual features.

Another challenge with using Word arises from file formats. Before 2007, versions of Microsoft Word used a proprietary binary format *(.doc)* to store Word files. While Word could export documents to many other formats, including plain text and HTML, the exporting process typically stripped away many formatting codes, effectively diminishing detail about the structure of documents. In the early 2000s, Microsoft began planning to use XML to store documents. The goal was to provide an open, documented standard for transparency and interoperability. The history of the development of what became the *Office Open XML standard (.docx)*[5] and its relationship to a previous competing XML-based standard, *Open Document Format* (ODF),[6] need not detain us here. While the move to these open formats certainly did advance the goals of interoperability, the hope that they would render the technical markup of word-processing documents legible to nonspecialists did not materialize. Few users, even among digital humanists, have dared to unpack these compressed files to inspect their arcane contents. The byzantine structure of these XML formats stymies data extraction from them.

In the end, how many authors care about the format of their electronic texts? The telos of WYSIWYG word processing is to produce a document that resembles the printed page.[7] Where XML-based formats behind the scenes have faltered in their original promise, the *Portable Document Format* (PDF) has solved the right side of the WYSIWYG equation. The PDF's success stems from its capacity to mimic the printed page. "Whether they render digitized text or text that has been born digital, as it were, pdf's present what are called page images," explains Lisa Gitelman. "They look something like pictures of pages produced by one printing process or another, or by word processing (Gittelman 2014, 115)." Given their success at this task, PDFs have become the central format of scholarly exchange, from Interlibrary Loan programs to open access repositories like *ArXiv* to "shadow libraries" such as *SciHub.* The technology backing implementations of the PDF standard belies its simplicity of presentation; the ISO standard for PDF 2.0 runs 986 pages.[8] Vanishingly few users tinker with the internals of the PDF format. Instead, they collect PDF documents in file folders, reference managers, and file-sharing servers, building their libraries of articles to read and cite.

The Digital Humanities community has taken different approaches to aligning digital writing with digital research. But breaking away from the spectral grasp of

5 ECMA-376 2021; see https://www.ecma-international.org/publications-and-standards/standards/ecma-376 (Accessed: 26 June 2024).

6 See https://www.oasis-open.org/2021/06/16/opendocument-v1-3-oasis-standard-published (Accessed: 26 June 2024).

7 As Gitelman (2014, 123) expresses the point, "For wysiwyg to work, there had to be continuity across screens (wys) and the pages printed out (wyg)."

8 See https://www.iso.org/standard/75839.html (Accessed: 26 June 2024).

paper-based formats proves challenging, if only because of disagreements about the path forward.

On the one hand, entrepreneurs have seized on the limitations of Microsoft Word to create niche products for authors and academics. First off the mark was *Nota Bene,* a word processor developed to format papers according to academic style guides such as the *Chicago Manual of Style* or the *American Psychological Association.* Steven Siebert, a doctoral candidate at Yale University studying philosophy and religion under Hans Frei, created *Note Bene* in the early 1980s out of frustration with the limitations of existing word processing software and has continued to market it successfully to academics for forty years.[9] More recently, *Scrivener*[10] has made headway among academics because of its nonlinear, associative approach to organizing long-form writing projects.

On the other hand, a series of alternatives to traditional word processors has emerged to render texts more tractable as data. An early alternative to the WYSIWYG paradigm was Donald Knuth's TeX (a play on τέχνη + X), which he developed in the late 1970s to facilitate the typesetting of his *The Art of Computer Programming.* In the 1980s, computer scientist Leslie Lamport developed macros to automate the features of TeX, making it more accessible to a broader group of users. "LaTeX is not a word processor!" notes the website for the LaTeX project. "Instead, LaTeX encourages authors not to worry too much about the appearance of their documents but to concentrate on getting the right content."[11] Of course, this gap between word processing and formatting with textual macros opened up a potential market. *Overleaf* has emerged as a productive compromise, providing a cloud-based authoring environment for LaTeX with conversion to PDF for visualization.[12] *Overleaf* also offers tools like collaborative editing and revision tracking. These days, a significant amount of scientific publishing takes place in LaTex; the American Mathematical Society "strongly encourage[s]" authors to use LaTeX because of its compatibility with its production systems and, presumably, its fine-grained handling of mathematical symbols.[13] LaTeX has made fewer inroads among digital humanists but may become more prominent as interdisciplinary work between theology and the sciences grows.

Markup languages push the distinction between content and appearance further by separating them nearly completely. By interweaving text and markup, markup languages make documents into data structures; unlike the .docx standard, the goal

9 In 1986, Kevin P. Roddy, a medievalist and early digital humanist (back when it was known as *Humanities Computing*) at the University of California, Davis, remarked, "Steven Siebert, the author of *Nota Bene,* had not at last report finished his dissertation in Philosophy at Yale. I hope that he has now returned to it, and left Version 3 to someone else. In America and elsewhere, we need as many philosophers as we do programmers (Roddy 1986, 95)."

10 See https://www.literatureandlatte.com/scrivener (Accessed: 26 June 2024).

11 See https://www.latex-project.org/about (Accessed: 26 June 2024).

12 See https://www.overleaf.com (Accessed: 26 June 2024).

13 See http://www.ams.org/publications/authors/tex/latexbenefits (Accessed: 26 June 2024).

of markup languages like *Text Encoding Initiative* (TEI) is to make XML documents equally human- and machine-readable. The TEI community has promoted the advantages of semi-structured formats for humanities since the 1980s; the commercial roots of markup languages like GML and SGML go back to the 1960s. In the late 1990s, the *eXtensible Markup Language* (XML) emerged as the dominant standard for creating markup languages.

Humanists have exercised an outsized influence on the XML community, co-developing the standards that emerged from SGML. The XML toolchain contains sophisticated tools for the entire lifecycle of documents, from definition to publication. The availability of open-source XML databases like *BaseX* and *eXist* has enabled scholars to publish their digital editions online. In contrast to relational databases, XML databases use *XQuery* as their combined application and query language.

Creating an XML-based digital edition has become significantly easier in recent years. *TEI Publisher* is a rapid application development environment built on *eXist* that provides the essential functionality for digital editions, including stylesheets for reading onscreen and in print, facetted search, and ready web hosting. *CETEIcean* takes a leaner approach, focusing on web publication of TEI documents using web components to integrate TEI elements into HTML.

XML and its related technology stack fell from favor in the 2010s. Just as XML reached a fever pitch of hype with industry leaders like Microsoft and IBM pivoting entire product lines to XML-based formats, the web community rebelled against XML. Douglas Crockford introduced JSON *(JavaScript Object Notation)* as a lightweight alternative to XML-based message-passing protocols like SOAP *(Simple Object Access Protocol).* JSON was not only more straightforward and, arguably, less verbose than XML, but it was also data-centric rather than document-centric.

The blowback against XML technologies in the early 2000s caused significant collateral damage in the Digital Humanities community. The community found itself isolated to some extent from the mainstream of technological development. XML-focused presentations began disappearing from the agendas of major tech events, though conferences like *Balisage* bound the XML community together.

To make XML more palatable to the workaday Digital Humanities scholar, *Performant Software Solutions* has released a TEI-based word processor called *FairCopy.*[14] *FairCopy* provides a graphic user interface for the TEI, allowing users to compose documents without having to wrangle angle brackets. Such graphic overlays over XML were already available in tools like *oXygen.* Still, *FairCopy* cleverly addresses the complexity of the TEI with its myriad elements and attributes by bundling relevant structures together. Still, you need to know some TEI to write effectively in *FairCopy.*

In the early 2000s, *Markdown* emerged as a stripped-down alternative to WYSIWYG systems and LaTeX and XML-based markup systems. John Gruber, the developer of *Markdown,* created the system out of frustration with online writing using

14 See https://faircopyeditor.com (Accessed: 26 June 2024).

HTML (Lockridge 2020). Flavors of *Markdown* have since evolved into the syntax of web-based writing, from *Discord* to *Github.* As Mailund (2019) remarks, "With *Markdown* you don't have quite the same power to control your formatting as you do in a language like LaTeX, but the simplicity of *Markdown* more than makes up for it."

But what good is *Markdown* for scholarly publishing in conventional contexts? After all, few scholarly publishers (with notable exceptions, such as *PubPub*[15]) accept manuscripts authored in *Markdown.* To facilitate its interoperability with other document formats, a command line program called *Pandoc* has become an essential companion to markdown in the scholarly writing workflow. John McFarlane, Professor of Philosophy at the University of California Berkeley, developed *Pandoc* and, not coincidentally, contributed to the standardization of *Markdown,*[16] to make switching between document formats straightforward.

The combination of *Markdown* and *Pandoc* has made it feasible to dispense with the WYSIWYG paradigm. As I type these words, for example, I am writing in *Visual Studio Code,* Microsoft's open-source code editor. My editing environment is more austere and componentized than Microsoft Word. A system of extensions allows me to add the features I want to the editor, including syntax highlighting for *Markdown,* a spellchecker, and a word counter. When I am finished writing, I use *Pandoc* to convert this document into Microsoft Word, PDF, or any required format. In a strange inversion, Microsoft Word .docx becomes an output file; in the final stage of the writing workflow, I will use *Pandoc* to convert the *Markdown* document(s) (along with accompanying references in BibTeX) into .docx for delivery to my editors.

4. Annotation

The problem of information overload is not a new issue for scholars. In *Too Much to Know,* Ann Blair explores pre-modern scholars' methods of organizing their research. Most techniques she surveys from the ancient and medieval world remain recognizable to contemporary researchers, though some have fallen out of favor. The act of marking up, taking notes, or otherwise annotating books has long served the practice of memory. The purpose of creating such annotations varied. "Annotations might make corrections to the text, add cross-references to similar material in the same or difficult texts, or include occasional words of praise or criticism," explains Blair (2011, 71), "but predominantly they flagged passages of interest […]." The tools have changed, but the motivations stay constant.

As a librarian, I have a professional duty to warn researchers against taking notes in borrowed books. The temptation to mark significant passages with light

15 See https://www.pubpub.org (Accessed: 26 June 2024).
16 See https://commonmark.org (Accessed: 26 June 2024).

pencil marks or, worse, a felt-tipped pen can prove strong. As a reader, I vacillate between abhorrence of texts defaced by yellow highlighting pen and amusement at witty comments or retorts sketched in the margins. In the contemporary world of digital scholarship, the restrictions on marking up documents have been lifted. When the bulk of your sources exist as PDFs or ebooks, analog forms of annotation no longer apply unless you print everything out. The developers of ebook software have sought to create digital analogs to marginal notes, allowing readers to mark up PDF documents with virtual highlights or put digital sticky notes in electronic books.

A new paradigm of social annotation has pushed beyond these simple replacements to instaurate marginalia for digital documents. Social annotation allows readers to mark up what they read online and, if they wish, to share their musings with private circles of others or the public at large. The W3C *(World Wide Web Consortium),* which creates standards for web-based protocols, has formulated a *Web Annotation Data Model* that specifies the conceptual underpinnings for web annotations. The definition of an annotation the working group provides is considerably more abstract than a pencil scrawl in the margins of a codex: "An annotation is considered to be a set of connected resources, typically including a body and target, and conveys that the body is related to the target."[17] This recondite language serves the ambitions of the web annotation; the goal is to provide a framework for the annotation of the entire web, including PDFs and audio/video resources. The motivating ideas behind web annotation harken back to the origins of the web. As Tim Berners-Lee envisioned the *Hypertext Transfer Protocol* (HTTP), the specification contained verbs to get web pages and post, put, and patch them. In other words, he imagined users not only reading and sharing information with web pages but updating them with new information. This utopian vision of a readable and writable web did not survive commercialization, but web annotation brings it back in revised form.

Commercial and open-source options now exist for sharing these digital marginalia. The leading open-source project is *Hypothesis.*[18] During the past decade, the *Mellon Foundation* provided grants to support the *Hypothesis* project and foster academic interest in web annotation. If you install the *hypothes.is* plugin in your browser, you can see where annotators have spoken back to web pages. Seeking annotations on the open web proves a hit-or-miss affair in practice. The stray comments you come across resemble graffiti rather than scholarly marginalia: On 16 October 2022, user jacobknight annotated the front page of Wikipedia with "GREEAATTTTT" and another user remarked a few days later in reply, "I know right?" While occasionally humorous, such scattered exchanges hardly constitute the thick web of scholarly annotation the fiscal sponsors had in view.

In contemporary humanities research, spreadsheets have become an essential companion to the word processor as the locus of note-keeping. The spreadsheet has

17 See https://www.w3.org/TR/annotation-model (Accessed: 26 June 2024).
18 See https://web.hypothes.is (Accessed: 26 June 2024).

taken over the role the box of note cards formerly played in scholarly research. Scholars in the humanities have become accustomed to populating spreadsheets with row after row of data as they read through ship manifests, colonial accounting books, or sermon texts. The spreadsheet serves as an ersatz database without the fuss of normalizing data. The ability to export from Microsoft Word or Google Sheets to *Comma Separated Values* (CSV), a loosely defined but ubiquitous format for sharing tabular data, makes moving data from spreadsheets to databases, code notebooks, and other data-driven software straightforward. Conferences like *csv,conf*[19] have emerged to facilitate the movement of data across these contexts. Once again, academics are converting office technologies to scholarly purposes rather than adopting unfamiliar and more pertinent tools such as web annotations.

5. Bibliography

Acknowledging and citing sources remains an essential element of academic discourse. A key mark of popular scholarship is its dispensation of a bibliographic apparatus. From academics crossing over to trade nonfiction, the transition can be jarring. In a tongue-in-cheek encomium titled *A Man without Footnotes,* Nathan Glazer remarks on what he terms Irving Kristol's *no footnotes* approach (Glazer 1995, 6)," a style that academics "consider a sign of arrogance (ibid.)" and "which has led to much agony among contributors, particularly since most of them are academics (ibid., 7)."

A significant focus of library instruction has been the teaching of bibliographic styles. At an undergraduate level, students trip over bibliographic rules through little fault of their own. "As college students go from class to class, they are often asked by their professors to use different citation styles, thereby preventing the students from becoming familiar with one particular style and all its nuances," writes Pfitzinger (2011, 28). "Since, for the foreseeable future, those in academia will continue to be forced to juggle multiple citation styles, familiarity with any one style manual will be difficult to achieve." An inventory of citation styles shows that more than ten thousand exist. However, many represent variations on the leading guidelines from the *American Philosophical Association* (APA), *Chicago Manual of Style, Modern Language Association* (MLA), and the *Institute of Electrical and Electronics Engineers* (IEEE).[20] Given the success of *Uniform Resource Locators* (URLs) for identifying webpages, we might imagine that a movement toward replacing bibliographic references with semantic identifiers would have occurred by now. While initiatives such as *Citation Identifiers* (CIDs), which condense bibliographic references into numeric facets, have

19 See https://csvconf.com (Accessed: 26 June 2024).
20 See https://citationstyles.org (Accessed: 26 June 2024).

gained traction in the sciences,[21] plans to replace bibliographic references with identifiers have faltered in the humanities.

What has sprung into the gap are bibliographic reference tools like *Zotero* (open source), *Endnote* and *Mendeley* (proprietary), along with a markup language called the *Citation Style Language,* which provides machine-readable descriptions of nearly all citation formats. Browser plug-ins allow readers to click on online articles and import the corresponding bibliographic data and the full text, if available, into their libraries. Using another plug-in, writers can use these citation managers to punctuate their manuscripts with dynamic links that generate the corresponding citations in the correct format on the fly. While maintaining a dynamic connection between the document and bibliographic manager proves handy when writing, authors must remember to *flatten* these references, that is, to break the links with their *Zotero* installation, before submitting their documents. Otherwise, their editors will receive a Word document with broken *Zotero* codes rather than a properly formatted bibliography.

6. Digital Publishing

To remark that the advent of the internet from the early 1970s and, in particular, the popularization of the *World Wide Web* from 1993 revolutionized scholarly communications is both a truism and easily overlooked. Before the advent of the *World Wide Web,* scholarly communications primarily took the form of sharing offprints. If your academic library did not subscribe to a journal, you could write to the author of the article, requesting an offprint be sent to you, or request a copy be sent to your library via interlibrary loan.

The desire to break away from the limitations of the print era has given rise to digital editions. Early in the era of personal computing, Ted Nelson explored the affordances of digital publication, dreaming of the *Xanadu Hypertext System.* “Its unique facilities of backtrack, linkage and windowing will allow the creation of new forms of multi-level, explorable collections and collages of material,” he opined, “without losing the well-defined authorship and ownership of all parts (Nelson 1981, chapt. 3,5).” Like Ted Nelson’s concept of hypertext, digital editions present scholarly information in nonlinear forms.

While the web’s openness allows freedom to experiment with nearly infinite varieties of forms with a decent knowledge of HTML and a pinch of *JavaScript,* scholars have gravitated toward web frameworks for good reason. Web frameworks not only provide the technical backbone of digital editions, connecting frontend user interfaces with backend databases but also encourage the development of third-party

21 See https://www.asme.org/publications-submissions/journals/administration/citation-identifiers-(cids) (Accessed: 26 June 2024).

ecosystems offering open source or commercial plugs-in or *skins.* The most popular frameworks among scholars are *WordPress* and *Drupal,* both open-source solutions that serve commercial audiences beyond the academy. In Digital Humanities circles, entrepreneurial scholars have responded to the mismatch of these web frameworks with scholarly ends by introducing academic content management systems, chief among them *Omeka*[22] and *Scalar.*[23] In the XML world, *TEI Publisher*[24] provides a *rapid application development* (RAD) environment for TEI texts. Suppose you have created a corpus of TEI documents using a customized ODD *(One Document Does It All),* that is, a specification that describes the semantics of your documents. In that case, *TEI Publisher* allows you to quickly create an online edition with the primary functions you would expect, including browsing, searching, and reading texts.

A counter-movement against database-backed content management systems has emerged in recent years under the rubric of *minimal computing.* By slicing through the complexity of web frameworks, *minimal computing* aims to streamline the online publication process and, at least in theory, to render the creation and consultation of digital editions more accessible. “While those who doubt their ability to learn how to code see the use of GUI-driven platforms as the key to access,” write Risam & Gil (2022, 16) in their preface to a special issue on *minimal computing,* “often these systems foreclose more control over the production of knowledge, and by extension, participation that is more meaningful to those who seek access.” Of course, they acknowledge the trade-offs when moving away from databases and graphic user interfaces (ibid.). Chief among those questions is that trading away user interfaces for code and data demands a higher level of computing knowledge.

The movement toward minimalism reaches its apex in the contemporary debate over data and interface. To what extent is an interface truly a requirement for sharing scholarly data on the web? In 2016, the Karl-Franzens-University in Graz hosted a conference titled *Digital Scholarly Editions as Interfaces.*[25] Dot Porter, Curator of Digital Research Services in the Schoenberg Institute for Manuscript Studies at the University of Pennsylvania and co-creator of the *OPenn* project, advocated prioritizing data over interface in digital editions. “Excellent, robust data with no interface isn’t easily usable (although a creative person will always find a way),” noted Porter in her keynote, “but an excellent interface with terrible data or no data at all is useless as anything other than a show piece.”[26] This call to prioritize data in digital editions has the potential to make academic projects far more interoperable, but is the academy ready to recognize these stripped-down digital editions as credible scholarly outputs?

22 See https://omeka.org (Accessed: 26 June 2024).
23 See https://scalar.me/anvc/scalar (Accessed: 26 June 2024).
24 See https://teipublisher.com (Accessed: 26 June 2024).
25 See https://informationsmodellierung.uni-graz.at/en/department/events/archive/digital-scholarly-editions-as-interfaces (Accessed: 26 June 2024).
26 See http://www.dotporterdigital.org/what-is-an-edition-anyway-my-keynote-for-the-digital-scholarly-editions-as-interfaces-conference-university-of-graz (Accessed: 26 June 2024).

The Digital Humanities community has pressed for recognition of formats beyond the article and the book as legitimate forms of scholarship. During the golden age of blogging in the aughts, scholars debated what weight to give blog writing in scholarly assessment and tenure review.[27] For the past decade, microblogging has taken over from blogging as an informal but essential running dialogue within the *Digital Humanities* research community. For better or worse, Twitter (now X) has served as the primary locus of this conversation. Ernesto Priego, Senior Lecturer in the Department of Computer Science at University College London, has analyzed tweets associated with the major Digital Humanities conferences since 2010.[28] As with many disciplines these days, following the right hashtags keeps you current with cutting-edge scholarship. But does it also promote a scholarly superficiality that favors, in the words of Lewis (2014), "the New New Thing?"

7. Literate Programming

The central idea behind literate programming is simple: optimize code for readability rather than conciseness. In contemporary Computational Humanities, literate programming frequently takes the form of code notebooks. A code notebook is a device that holds explanatory narrative and executable code together. A code notebook offers two kinds of cells: code cells and text cells. The code cells provide an environment for executing code and, if relevant, for displaying the output of the computation. The text cells, by contrast, contain non-executable information, generally a narrative about how the code blocks function or the purpose of their computation.

The most popular form of code notebook is called a *Jupyter notebook*. Since these notebooks were developed in the *Python* programming community,[29] *Python* is the default choice of kernel, or programming environment, for *Jupyter*. But *Jupyter* is fully extensible, and kernels exist for many other languages, including *R* and *SQL,* among many others. Google has popularized *Jupyter* by offering its *Colaboratory* (or *"Colab"*) service;[30] *Colab,* which offers both free and paid service, connects to cloud servers on the *Google Cloud Platform* (GCP), eliminating the need to set up hosting either locally or remotely for *Jupyter. Colab* has proved immensely popular among artificial intelligence specialists. These days, the "Open in *Colab*" button regularly appears alongside

27 See https://www.science.org/content/article/science-blogging-and-tenure (Accessed: 26 June 2024).
28 See https://ernestopriego.com/2019/07/15/dh2018-and-dh2019-twitter-archive-counts-a-comparison (Accessed: 26 June 2024).
29 See https://cs.lbl.gov/news-media/news/2021/project-jupyter-a-computer-code-that-transformed-science (Accessed: 26 June 2024).
30 See https://colab.research.google.com (Accessed: 26 June 2024).

papers, encouraging readers to give the code a spin and try their new machine-learning techniques for themselves.

Not all code notebooks work the same way. *Observable,* for instance, uses *JavaScript* rather than *Python* as its primary coding language.[31] Whereas *Jupyter notebooks* execute code sequentially from top to bottom, *Observable notebooks* execute code topologically; that is, code cells update whenever any of the cells they depend on change.[32] Technically, this paradigm is termed functional reactive programming, but the way *Observable* works in practice is similar to a spreadsheet. For example, if you change a cell in a spreadsheet, your calculations that depend on that cell will update accordingly. *Observable* works similarly, responding to updates in both code and data dynamically. Given digital humanists' ubiquitous use of spreadsheets, the *Observable* paradigm may feel more familiar than the top-to-bottom execution style of *Jupyter notebooks.*

The move toward literate programming with code notebooks has not gone without criticism. While notebooks provide an ideal environment for exploratory data analysis, they suffer from bloat and complexity as they increase in size. From a pedagogical standpoint, students may pick up problematic coding habits from notebooks (Johnson 2020). But, despite the criticisms, code notebooks have found their place as essential complements to scholarly papers in computer science and specific sectors of the Digital Humanities.

8. Quantification of Theological Knowledge

The transformation from analog to digital research methods inevitably disrupts our approach to scholarly analysis. To play on Wittgenstein's adage, Digital Humanities "does not leave everything as it is." The shift to computational methods in theology carries along a host of other techniques, including programming, data management, and statistical modeling. A trope about data science is that it exists at the intersection of computer science, statistics, and disciplinary knowledge. The problem, to return to Daniel Allington, is that almost nothing of computer science or quantitative methods is taught during seminary or Ph.D. programs in the humanities, leading to his sweeping judgment, "there is scant prospect for the development of quantitative methods in primarily hermeneutic disciplines such as literature, philosophy, or theology (Allington 2022, 381)." As a generalization, this is essentially correct, but there are exceptions, both at the curricular and the (sub-)disciplinary level.

What is the solution to this lack of background knowledge among theologians and scholars of religious studies?

31 See https://observablehq.com (Accessed: 26 June 2024).

32 See https://observablehq.com/observablehq/how-observable-runs (Accessed: 26 June 2024).

On the one hand, programs like *Software Carpentry*[33] and *Data Carpentry*[34] have emerged as boot camps to bring graduate students and scholars up to speed with data-intensive programming. These carpentries provide crash courses on *git,* the *Unix shell, Python,* and *R* programming. In Digital Humanities circles, intensive summer programs, like the *Digital Humanities Summer Institute* (DHSI) at the University of Victoria, have trained cohorts since 2001 in techniques ranging from textual encoding to network analysis. These boot camps evince mixed success. While they orient newcomers to the toolsets of the Digital Humanities, they exude a similar misleading appeal as language courses pitched to business travelers: promising functional proficiency after only a couple of weeks of study. Your actual mileage may vary. On the statistics side, there are fewer opportunities to develop fluency. Arnold & Tilton (2019, 293) have purported to "show how statistics – the organization, analysis, interpretation, and presentation of data – is a fundamental interlocutor and methodological approach for the Digital Humanities." But who will teach faculty the requisite mathematics and train them in data analysis?

The solution to this quandary is generally not for religious studies scholars to develop computer science and statistics expertise. The path toward gaining facility in these disciplines is long and inevitably involves passage through mathematical fields like calculus and linear algebra. It is safe to say that we will not see these subjects appearing in seminary curricula in the foreseeable future.

> No one person can be expected to understand such a wide range of disciplines brought together in the digital humanities in the depth required to develop the innovative insights and methods that are the promise of the field. Rather, the digital humanities should welcome statistics to the table, and this begins by better acknowledging the critical role of statistics in the field (ibid., 298).

The same goes for software engineers, project managers, metadata specialists, and DevOps experts. Building a Digital Humanities project requires a team or, to use the nomenclature of the natural sciences, a lab. We see movements in this direction in the Digital Humanities, but the organizational apparatus remains challenging to muster. As we shift from conventional research environments to Computational Humanities (and theology), the emergence of these labs may be the most tangible sign of the transformed research environment.

33 See https://software-carpentry.org (Accessed: 26 June 2024).
34 See https://datacarpentry.org (Accessed: 26 June 2024).

9. From Saturation to Datafication

In this essay, we surveyed a series of digital surrogates for analog research and publication methods. While we have covered much territory, we've only just explored the landscape. If you are feeling overwhelmed by the options, you are not alone. But how do digital humanists deal with media saturation?

The shift in emphasis from interface to data points toward one path forward. In *On the Existence of Digital Objects,* Yan Hui explores the being of digital objects in dialogue with Husserlian phenomenology and post-Husserlian ontology. As he remarks near the outset of his study, "the term 'digital object' remains ambiguous here, because the vast quantity of digital objects are comparable in breadth and diversity to the vast array of animal species (Hui 2016, 48)." We interact with increasing numbers of digital objects, running the gamut from Instagram posts, tweets, Google Docs, file folders, Android or iPhone apps, and so on. These objects multiply every time a new medium comes into existence, inflating our digital ontologies. Putting the point more familiarly, whenever we download a new app on our phones or try out a new educational technology, we need to familiarize ourselves with its way of cutting up its digital domain. As the channels proliferate, we tend to experience a diminution of interest. "How do we share a post in Mastodon again? How can I change the order of my feed in Threads?" In the language of media studies, our digital media environment has become saturated; adding novel technologies to our academic workflows threatens to sap energy rather than accelerate our scholarship.

In the Digital Humanities, scholars have responded to the saturation of our academic media landscape by moving away from digital interfaces toward data, metadata, and code. There remains tremendous diversity at this level of scholarly computing as well. But, as Yuk Hui suggests, the operating ontologies of Computational Humanities – and, by extension, Computational Theology – become perspicuous when we strip away the interfaces and examine the code, data, and metadata that animate them behind the scenes (ibid., 26). Still, the shift to the command line or code notebook threatens to exclude scholars who lack the time, resources, or inclination to explore the arcane languages and protocols that animate it. As a remedy to media saturation, this kind of computational reductionism works, but it may also narrow the field. How many digital humanists have the capacity, or inclination, to retool completely their scholarship around the software engineering toolchain?

In the near term, most theological scholars will simply pick and choose academic tools from conventional (analog) and new (digital) media. They will continue to borrow books from the library while downloading PDFs online. They will write marginal comments in their texts while keeping track of key dates in Excel documents. And they will publish monographs with university presses while showcasing their projects in multimedia digital editions. In other words, the digital will not displace the analog but supplement it. The question is whether this mixed media approach will persist. Will we look back on this era with nostalgia for the books, pencils, file cards, and notebooks that have since vanished from our scholarly environment? Or will we find some efficient equilibrium that blends the best of analog and digital into a perduring scholarly synthesis?

References

Allington, D. (2022). The Place of Computation in the Study of Culture. In J. O'Sullivan (Ed.), *The Bloomsbury Handbook to the Digital Humanities* (pp. 373–384). London: Bloomsbury Publishing.

Arnold, T., & Tilton, L. (2019). New Data? The Role of Statistics in DH. In M.K. Gold & L. Klein (Eds.), *Debates in the Digital Humanities* (pp. 293–299). Minneapolis: University of Minnesota Press.

Blair, A.M. (2011) *Too Much to Know. Managing Scholarly Information Before the Modern Age.* New Haven: Yale University Press.

Bryant, J., & Fondren, W. (2009). Displacement Effects. In R.L. Nabi & M.B. Oliver (Eds.), *The Sage Handbook of Media Processes and Effects* (pp. 505–517). Los Angeles: Sage.

Clark, A., & Chalmers, D. (1998). The Extended Mind, *Analysis*, 58(1), 7–19. DOI: https://doi.org/10.1093/analys/58.1.7 (Accessed: 26 June 2024).

Gitelman, L. (2014). *Paper Knowledge. Toward a Media History of Documents.* Durham: Duke University Press Books.

Glazer, N. (1995). A Man Without Footnotes. In Ch. DeMuth & W. Kristol (Eds.), *Neoconservative Imagination. Essays in Honor of Irving Kristol* (pp. 3–12). Washington, D.C: AEI Press.

Hassan, R. (2023). *Analog.* Cambridge: The MIT Press [= *The MIT Press Essential Knowledge Series*]. DOI: https://doi.org/10.7551/mitpress/13644.001.0001 (Accessed: 26 June 2024).

Hui, Y. (2016). *On the Existence of Digital Objects.* Minneapolis: University of Minnesota Press.

James, R. (2010). An Assessment of the Legibility of Google Books, *Journal of Access Services,* 7(4), 223–228. DOI: https://doi.org/10.1080/15367967.2010.503486 (Accessed: 26 June 2024).

Johnson, J.W. (2020). Benefits and Pitfalls of Jupyter Notebooks in the Classroom. In *Proceedings of the 21st Annual Conference on Information Technology Education* (pp. 32–37). New York: Association for Computing Machinery. DOI: https://doi.org/10.1145/3368308.3415397 (Accessed: 26 June 2024).

Kirschenbaum, M.G. (2016). *Track Changes. A Literary History of Word Processing.* Cambridge: Belknap Press.

Lewis, M. (2014). *The New New Thing. A Silicon Valley Story.* New York: W.W. Norton & Company.

Lockridge, T. (2020). *Writing Workflows. Beyond Word Processing.* Michigan: University of Michigan Press.

Mailund, Th. (2019). *Introducing Markdown and Pandoc. Using Markup Language and Document Converter.* Berkeley, CA: Apress. DOI: https://doi.org/10.1007/978-1-4842-5149-2 (Accessed: 26 June 2024).

Marcum, D., & Schonfeld, R.C. (2021). *Along Came Google. A History of Library Digitization.* Princeton: Princeton University Press.

Nelson, T. (1981). *Literary Machines.* Sausalito: Mindful Press.

Newell, J., Pilotta, J.J., & Thomas, J.C. (2008). Mass Media Displacement and Saturation, *International Journal on Media Management,* 10(4), 131–138. DOI: https://doi.org/10.1080/14241270802426600 (Accessed: 26 June 2024).

Pfitzinger, S.G. (2011). The UPs and Downs of Citation Titles. What Librarians Should Know When Helping Students, *Indiana Libraries,* 30(2), 25–29.

Risam, R., & Gil, A. (2022). Introduction. The Questions of Minimal Computing, *Digital Humanities Quarterly,* 16(2), 1–34. URL: http://www.digitalhumanities.org/dhq/vol/16/2/000646/000646.html (Accessed: 26 June 2024).

Roddy, K.P. (1986). Once More into the Breech. Computer Literacy and the Humanities, *Historical Social Research,* 40, 91–95. URL: https://www.jstor.org/stable/20755058 (Accessed: 26 June 2024).

Schatz, B.R. (1997). Information Retrieval in Digital Libraries. Bringing Search to the Net, *Science,* 275(5298), 327–334. DOI: https://doi.org/10.1126/science.275.5298.327 (Accessed: 26 June 2024).

Wasiak, J. (2008). Epistemological Implications of Media Saturation, *The International Journal of Technology, Knowledge, and Society,* 4(2), 113–118. DOI: https://doi.org/10.18848/1832-3669/CGP/v04i02/55862 (Accessed: 26 June 2024).

Glossary*

Kevin Wunsch[a] and Christopher A. Nunn[b]

[a] https://orcid.org/0000-0003-1491-747X, [b] https://orcid.org/0000-0001-7208-8636

Algorithm
An algorithm is a step-by-step, predetermined procedure for solving a specific task. If the entry data remains unchanged, the algorithm will always produce the same result. In digital editing, the *I-P-O* principle is used (= **I**nput, **P**rocessing, **O**utput): the XML file is processed, and the output is an html-structure.

Alignment
Alignment is a term used in Bioinformatics where similarities within biological sequences are recorded. In the Digital Humanities, the term refers to the comparison of data from different sources. Alignment makes the complex correlations visible.

Allographs
Allographs are variants of a sign that represent the same linguistic value. Allographs often occur in different writing systems. The differences are often small and do not have any effects on the meaning.

APC (Article Processing Charge)
APC refers to the cost paid to the publisher when one wants to have a work published in Open Access.

API
An API is an interface used to interact with the data of a website. *Information retrieval* usually occurs through interfaces. Common standards for interfaces on the web are REST *(Representational State Transfer)* and SOAP *(Simple Object Access Protocol).*

Artifact
An artifact is a historical remnant that can be the subject of research in some form.

Augmented Reality
The computer assisted overlay and enhancement of reality with one or more layers of artificial content that extend and enrich the real world.

* This chapter was translated from German by Kevin Wunsch.

Kevin Wunsch and Christopher A. Nunn: Glossary, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 463–476. DOI: https://doi.org/10.11588/heibooks.1521.c21965

Backend
The backend is the invisible part of a software system that is responsible for processing and storing data.

Boxplot
A boxplot is a visual summary of variability. In addition to the median, a boxplot shows quartiles, minimum and maximum values, and prominent outliers. Fig. 1 at end of glossary on p. 475, for example, presents 2000 newspaper articles from *DIE ZEIT* published between 2001 and 2014. The boxplots illustrate the average text length of 200 documents, which were categorized into different classes by the *ZEIT* editorial team.

Close Reading
Franco Moretti sharply contrasts Close Reading with Distant Reading. Close Reading is a method of text analysis in which a text is carefully examined in detail to understand its formal and linguistic features, as well as its various layers of meaning. This method is considered a conceptual precursor to *New Criticism,* a literary theory movement that utilizes formalized Close Reading as a central analytical technique.[1]

Codec
A portmanteau of the words *Coder* and *Decoder.* A codec describes a pair of algorithms that encode and decode data, often for the purpose of more efficient data transmission.

Container
Container systems allow developers to scale their software more easily. One widely used container system is the open-source solution *Docker.*[2] With the help of container systems, the problem of "it works on my machine" is mitigated, as the development and deployment environments are identical.

Content Management System (CMS)
A software (collection) that can be used to manage, edit, and publish content in the digital realm, even without technical knowledge. *WordPress* is a widely used CMS.

Crowdsourcing
Crowdsourcing refers to the outsourcing of resources to the community or *crowd.* For example, archive material can be transcribed by volunteers (e.g., in the commu-

1 Cf. Moretti, F. (2000). Conjectures on World Literature, *New Left Review,* 1, 54–68; Herrnstein-Smith, B. (2016). What was "Close Reading?" A Century of Method in Literary Studies, *Minnesota Review,* 87, 57–75.
2 See https://www.docker.com (accessed on 14 July 2024).

nity project *Consilium Communis* for the transcription of historical documents in the Neuss city archive).[3]

Data
Data is the digital representation of information from various research fields. This can include be texts, pictures, audio recordings, and videos, as well as metadata or aggregated data. A basic distinction is made between structured and unstructured data.

Data (structured)
A good example for structured data is XML data, which maps the hierarchical structure of the text.

Data (unstructured – messy/fuzzy)
Unstructured or *messy/fuzzy* data are incomplete or inconsistent. They are generally more difficult to interpret and analyze than structured data. The algorithmic evaluation of unstructured data is also more complex than structured data.

Deep Learning
Deep Learning is a branch of *Machine Learning* based on neural networks. These multi-layered networks are optimized for the processing of large amounts of data and can automatically recognize patterns and features in data.

Diasystem
A diasystem refers to a system in which variants and varieties of a language (including dialects) occur in a defined area, such as geographical or social contexts. The varieties can cover all areas of linguistics. In linguistics, diasystems are used to analyze the diversity within a language.

Digital Turn
The term describes the shift from traditional methods and practices to the widespread adoption of digital technologies and processes. The *digital turn* has revolutionized several academic fields.

Disintermediation
Disintermediation can also be described as “cutting out the middleman.” The term indicates the avoidance and bypassing of classic mediators.

Distant Reading
Distant Reading involves analyzing texts *from a distance.* Algorithms and computer-supported methods enable the examination of large corpora of data (big data).

3 See https://www.stadtarchiv-neuss.de/nachrichten-detail/198.html (accessed on 14 July 2024).

While Close Reading concentrates on individual texts, Distant Reading often uses quantitative approaches, text mining or pattern analysis, to make claims about a larger datasets.

Entity
In computer science, an entity refers to objects or things that can be described by data. They can be real or abstract. In the Digital Humanities, there are also several vocabularies for describing objects. The CIDOC-CRM[4] standard defines an entity as a thing and entities as things like places, persons, or documents.

Feature
A feature is a specific characteristic of an object (or entity), such as an artifact or a text.

Figure Idiolect
A figure idiolect refers to the individual way of speaking of a literary figure, thus making the figure unmistakable and emphasizing the figure's personality and socio-cultural background.

Framework
Broadly speaking, a framework is a defined collection of reusable (code) building blocks that simplify software development. Frameworks are essential as the basic framework for many areas, since they not only simplify and accelerate the development process, but also allow developers to focus on the specifics.

Frontend
The *visible* parts of a website or application are the frontend. Popular frameworks for frontend web development are *ReactJS*[5] or *Angular*[6].

Geocoding
Geocoding refers to the encoding of geographic coordinates so that the locations can be read by computers. It enables the visualization of location data on maps and in geographic information systems. The most well-known tools for processing this data are *ArcGis*[7] and the Open Source tool *QGIS*[8].

GLAM (Galleries, Libraries, Archives, and Museums)
Generally, all institutions of cultural heritage fall within this broad category.

4 See https://www.cidoc-crm.org (accessed on 14 July 2024).
5 See https://react.dev (accessed on 14 July 2024).
6 See https://angular.dev (accessed on 14 July 2024).
7 See https://www.arcgis.com/index.html (accessed on 14 July 2024).
8 See https://www.qgis.org (accessed on 14 July 2024).

Graph Database

A graph database saves data in nodes (entities) and edges (relationships). These databases enable the modelling and querying of complex relational structures.[9] In addition to XML databases, graph databases are a common type of database for creating digital editions. However, XML can also be viewed as a directed graph whose elements and attributes represent nodes, while the directed edges reflect the hierarchical relationships.

GUI (Graphical User Interface)

GUI refers to a graphical user interface, probably the most common, but not always the most efficient way to interact with an application.[10]

Heat Map

Data is highlighted in color in a heat map, e. g., the positions of a soccer player during a game.[11]

(HMM) Hidden Markov Model

Hidden Markov models are often used in linguistics or bioinformatics to recognize patterns. Latent (or hidden) states are inferred from a series of observable events and data points. The model consists of states, state transition possibilities, and output possibilities. They describe the likelihood of a certain observable event occurring.

HTML (Hypertext Markup Language)

HTML is the markup language of the internet, i. e., the descriptive language that makes up web pages.

Inference Server

An inference server is a specialized server supported by a GPU *(Graphics Processing Unit)* that can make predictions as quickly as possible – ideally, almost in real time – with the help of machine learning models.

Information Retrieval

Typically, information retrieval refers to the algorithm-driven process of identifying and retrieving data that matches a user's search query.

9 Cf. Kuczera, A. (2017). Graphentechnologien in den Digitalen Geisteswissenschaften, *ABI Technik,* 179–196. https://doi.org/10.1515/abitech-2017-0042 (accessed on 14 July 2024).

10 Cf. Drucker, J. (2011). Humanities Approaches to Graphical Display, *digital humanities quarterly,* 5(1), 1–52. http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html (accessed on 14 July 2024).

11 On uses of heatmaps, see, for example, Memmert, D., & Raabe, D. (eds.). (2019). Revolution im Profifußball. Mit Big Data zur Spielanalyse 4.0. 2nd ed. (pp. 92 f.). Berlin: Springer. https://doi.org/10.1007/978-3-662-59218-2 (accessed on 14 July 2024).

Java
Java is a programming language on which many database applications are based. *ExistDB*[12], the *standard tool* for creating digital editions, or *ediarum*[13], a framework for creating digital editions, are both written in Java.

JavaScript (JS)
JavaScript has long been known as a language for manipulating websites, such as changing the color of a button when clicked. It has been increasingly used in backend applications. Despite the name, this programming language is not related to Java.

Jupyter Notebook
Jupyter Notebook is an interactive computing platform that allows users to collect all aspects of work in a web document. It has become the standard for the development and presentation of results, particularly in the fields of *data science* and *data driven* sciences.

Kernel
The kernel is a core component of Unix-based operating systems *(Linux, MacOS).* It contains basic information and operations for the computer. The kernel is effectively the heart of the computer and has complete control over everything in the system.

Lemmatization
Lemmatization is the process of reducing a word to its base form, i. e., to their *lemma,* such as *went* → *to go.*

Linked (Open) Data
Linked Open Data is a part of the semantic web, which can be described as the *internet for machines* and consists of several levels. The semantic web is not about making data and information available for human processing, but about enabling *intelligent machines* to be able to use the data. This is achieved through linked data.

Machine Learning
Machine learning, a branch of artificial intelligence, enables computers to learn from data and perform tasks without explicit programming. It uses algorithms to extract predictions and recognize patterns from data. A distinction is made between *supervised learning* and *unsupervised learning.* In supervised learning, algorithms are trained using labeled data, while in unsupervised learning, algorithms are used to recognize patterns and structures in unlabeled data. In *reinforcement learning,* both negative and positive results are evaluated, such as winning a game, so that similar

12 See http://exist-db.org/exist/apps/homepage/index.html (accessed on 14 July 2024).
13 See https://www.ediarum.org (accessed on 14 July 2024).

strategies for problem solving are used more frequently, while less efficient ones are used less often.

Macroanalysis
Macroanalysis refers to the approach of analyzing large text corpora using quantitative methods (such as distant reading). This term was popularized by Matthew Jockers' 2013 book *Macroanalysis.*

Mel Frequency Cepstral Coefficients (MFCC)
These coefficients are one of the central features in computer-aided audio processing. They provide an efficient representation of the essential characteristics of audio signals.

Mid-Range-Reading
Methodologically, mid-range reading falls between close reading and distant reading.[14] In German-speaking countries, the term *scalable reading* is more common, a reading method that alternates between close and distant reading depending on the subject matter and the research question.[15]

Mixed Methods
Methods from different scientific traditions are combined – a common example is the simultaneous use of qualitative and quantitative methods.

Neural Network
This biological term describes a group of interconnected units in the brain called neurons. In information science, however, a neural network consists of various information processing units in a layered structure. They weight inputs to make predictions based on patterns.

N-Gram
An n-gram is used to describe coherent sequences from a text: "here is" is a bigram, while "here is text" is a trigram. With the help of n-grams, patterns can be recognized, or contexts and frequencies of word chains can be analyzed.

14 Cf. Booth, A. (2020). Mid Range Reading. Not a Manifesto, *Publications of the Modern Language Association of America,* 132(3), 620–627. https://doi.org/10.1632/pmla.2017.132.3.620 (accessed on 14 July 2024).

15 See, for example, Krautter, B. (2024). The Scales of (Computational) Literary Studies. Martin Mueller's Concept of Scalable Reading in Theory and Practice. In F. Armaselu & A. Fickers (eds.), *Zoomland. Exploring Scale in Digital History and Humanities* (pp. 261–286, esp. p. 262). Berlin/Boston: De Gruyter Oldenbourg [= *Studies in Digital History and Hermeneutics,* 7].

Natural Language Processing (NLP)
NLP is a subfield of artificial intelligence that focuses on the interaction between computers and human language, encompassing methods for processing, analyzing, and generating natural language. Applications range from machine-assisted translation and text classification to language generation and sentiment analysis. The methods of linguistics and computer science are combined to enable computers to understand human language.

Noise
Noise within a data set disturbs and interrupts patterns. These are irrelevant or random data variations that can negatively impact the performance of models. Sometimes it is difficult to distinguish important information from noise. Various methods of data cleansing can be helpful.

OJS (Open Journal Systems)
OJS refers to open source software for managing and publishing scientific journals, primarily used in Open Access publications. For example, the *Zeitschrift für digitale Geisteswissenschaften*[16] is based on OJS.

Operationalization
Operationalization refers to the process of converting abstract constructs into variables, thus enabling machine studies by defining quantifiable indicators.[17]

Optical Character Recognition (OCR)
OCR is used for the computer-generated recognition of text. Various tools are available for this purpose, some of which operate on personal devices while others are web-based. *Transkribus*[18], *OCR-D*[19], *Tesseract*[20], and *Abby FineReader*[21] are some of the most well-known tools.

16 See https://zfdg.de (accessed on 14 July 2024).
17 For a detailed description of this technique, see Krautter, B., Pichler, A., & Reiter, N. (2023). Operationalisierung. In AG Digital Humanities Theorie des Verbandes Digital Humanities im deutschsprachigen Raum (ed.), *Begriffe der Digital Humanities. Ein diskursives Glossar* (no pag.). Wolfenbüttel: Herzog August Bibliothek [= *Zeitschrift für digitale Geisteswissenschaften. Working Papers,* 2]. https://doi.org/10.17175/wp_2023_010 (accessed on 14 July 2024).
18 See https://www.transkribus.org/de (accessed on 14 July 2024).
19 See https://ocr-d.de (accessed on 14 July 2024).
20 See https://github.com/tesseract-ocr/tesseract (accessed on 14 July 2024).
21 See https://pdf.abbyy.com (accessed on 14 July 2024).

Overfitting
Overfitting is the effect of an *over-trained* model in machine learning. The model has internalized the training data – including exceptions and noise – so thoroughly that its ability to generalize to new data deteriorats. This means the model performs very well on the training data, but poorly on the test data.

Part-of-Speech-Tagging (POS Tagging)
POS tagging is the process of labeling the different types of speech in a text. Doing so helps one understand the structure of the text, which is a fundamental step in many NLP tasks.

PID (Persistent Identifier)
Persistent identifiers are permanently available and immutable. The reference operates through one of several systems, the best known being PURL *(Persistent Uniform Resource Locator),* DOI *(Digital Object Identifier),* and URI *(Uniform Resource Identifier).*

Pipeline (NLP)
A pipeline is a sequence of steps for text processing or analysis. A possible pipeline would be, for example, tokenization → POS tagging → feature extraction, which structures a raw text that can then be used for machine analyses.

Plugin
A plugin is an optional extension to an application. Examples include ad blockers for browsers or the citation plugin *Zotero* in Office applications.

React
React is a widely used frontend framework based on *JavaScript.* It accelerates the development of responsive and easily maintainable web applications, often benefiting accessibility.

Relational Databank
A relational database is a system that stores and manages data in tables. The relationships between the tables are defined by keys, and queries are executed using a form of SQL.

Retrodigitization
The digitization of analogue media, such as books, photographs, or audio recordings, is known as retrodigitization. This process makes the media accessible, searchable, and available for the long term.

Scatterplot
A scatterplot is a graphical representation of two variables and their value pairs in a two-dimensional coordinate system. The goal is to highlight relationships between the variables and to make possible connections visible. E.g., Christof Schöch examined 60 themes and their occurences in 890 French dramas between 1610 and 1810 (from the Classical period to the Enlightenment, based on Paul Fièvre's Théâtre classique collection). He visualized the results in a scatter plot (see Fig. 2 at end of glossary on p. 475).

Segmentation (of texts)
Segmentation refers to the division of a text into meaningful smaller units, such as sentences, paragraphs, or individual words. The segmentation of a text often enhances the possibilities for computer-aassisted analysis.

SIFT Methods (Scale Invariant Feature Transform)
A SIFT method is an algorithm from image processing that recognizes and describes central image elements. The method is used to compare images.

Skin
A skin is a design or appearance that can be customized by users that does not affect the functionality.

Softmax Function
The Softmax function is used particularly in neural networks to generate probability distributions over a set of classes. it transforms vectors into probability distributions so that the sum of all probabilities equals 1.

SQL (Structured Query Language)
SQL is the foundational language of relational database systems. Large parts of the internet use relational databases.

Stemming
Stemming is a method in NLP that aims to reduce different forms of a word to the common stem. Suffixes and prefixes are removed so that only the stem of the word remains (e.g., going → go). Stemming is an important method for classification questions, such as sentiment analysis.

Stop Words
Stop words are words that are considered to have litte semantic significance. Thus, stop words are typically removed to improve the relevance and efficiency of analyses and queries. Depending on the genre of the text, it is important not to rely on automated stop word lists, but rather to develop them "professionally and appropriately."

Otherwise, key content may not be machine analyzed (consider Hamlet's "To be or not to be").[22]

TEI (Text Encoding Initiative)
TEI is the standard for encoding texts according to a defined vocabulary that can be continuously expanded by the community and adapted to changing conditions. It captures content and structural units in texts using tags, making them analyzable. Due to various XML dialects, the TEI guidelines are also tailored for specific areas of application, e.g., *TEI EpiDoc* for the structured markup of scientific editions of ancient documents, particularly inscriptions and papyri, or *TEI Correspondence* for marking up letters.

Tokenization
Tokenization is the process of dividing text into defined unites, typically words or phrases. This enables a variety of computer-aided analysis methods.

Toolchain
A toolchain is a collection of tools used in a consistent order to perform complex developmental tasks. Each specialized tool adresseses a particular set of functions. Together, these tools enable the development process.

Transformer
Developed in 2017, transformers are a relatively new concept in machine learning. The powerful *Large Language Models* (such as BERT or Chat-GPT) would not be possible without transformer models. These models are distinguished from other models by their self-attention mechanisms, parallel processing, encoder-decoder architecture, position coding, and high scalability.

Treebank
Treebanks were introduced by Geoffrey Leech in the 1980s as a concept and method for capturing sentence structure in a tree structure. The tree divides the sentence into its components, such as subject – object – verb.[23]

Type-Token Ratio
The type-token Ratio is a measure of the lexical diversity of a text. It describes the ratio of unique words *(types)* to the total number of words *(tokens).* The higher the

22 Cf. Schubert, Ch. (2021). Digital Humanities auf dem Weg zu einer Wissenschaftsmethodik. Transparenz und Fehlerkultur, *Digital Classics Online*, 7, 39–53, here: 41–42. https://doi.org/10.11588/dco.2021.7.82371 (accessed on 14 July 2024).

23 Cf. Leech, G., & Eyes, E. (1997). Syntactic Annotation. Treebanks. In R. Garside, G. Leech & T. McEnery (eds.), *Corpus Annotation* (pp. 34–52). New York: Addison Wesley Longman.

value, the more diverse the vocabulary, while a lower value indicates a very limited vocabulary.

Unicode (Universal Character Encoding)
Unlike other standard sets of characters, Unicode is not focused on a subsystem of human languages, but aims to represent all characters in the binary system comprehensively. It also serves as the technical foundation for character encoding.

Unix
Unix was an operating system developed in the 1960s and 1970s. It gave rise to several well-known and widely used operating systems – the most notable are *Linux, BSD* and *macOS.*

XML (Extensible Markup Language)
XML is a markup language used to represent structured data. Since its first publication, XML has been widely adopted in many areas of digital life, including the digital editing community. TEI-coded texts are usually marked up in XML.

Vector
A vector is a mathematical quantity covering an n-dimensional space. Vectors are often used in algebra, physics, or computer science. In NLP, a vector represents the dimensions of a word. Cities are often cited as an example: The cities of Berlin and Paris are each capitals of the countries in which they are located (Germany, France), but their dimensions, such aspopulation size and spoken language, for example, differ.

Virtual Reality
In contrast to augmented reality, virtual reality is an artificial reality. Currently VR headsets, such as the *Oculus Rift,* are widely used, at least for computer and console games. The simplest virtual reality is therefore a computer game that transports you into a fictional world. However, the use of virtual reality for digital research is also gaining traction.[24]

Visual Computing
Visual computing describes an interdisciplinary field of computer science that encompasses all aspects of digital work with images. This includes the generation of images and the automated evaluation of image information *(Computer Vision),* as well as image processing and visualization.

24 See, for example, the *VR-Lab* at the Institute for Digital Humanities at the Georg-August-Universität Göttingen, https://www.uni-goettingen.de/de/vr-lab/662748.html (accessed on 14 July 2024). Here, “virtual, augmented, and mixed reality technologies are used to reconstruct spaces of the past and critically question these visualizations.”

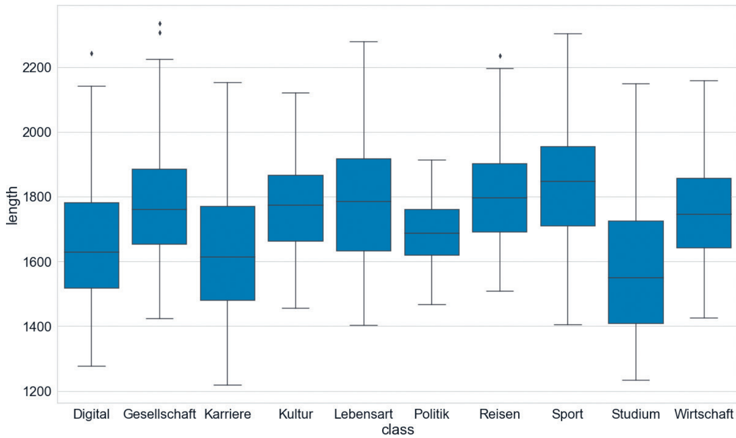


Fig. 1 Distribution of text length in the ten classes of newspaper articles.

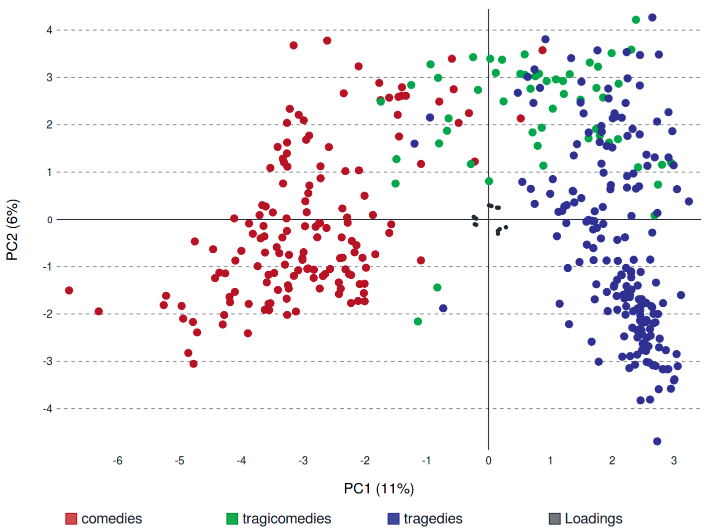


Fig. 2 Scatter plot after a Principal Component Analysis (a method of multivariate statistics) of French dramas and the likelihood of the occurrence of 60 themes within them.

Figure Credits

Fig. 1: Du, K. (2022). *Zum Verständnis des LDA Topic Modeling. Eine Evaluation aus Sicht der Digital Humanities* [PhD Thesis]. Online: Universität Würzburg. Here: p. 74, fig. 5.1 (CC BY 4.0). https://opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/34826/file/Du_Keli_Dissertation.pdf (Accessed: 14 July 2024).

Fig. 2: Schöch, Ch. (2017). Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama, *digital humanities quarterly,* 11(2), 1–53. Here: p. 35, fig. 10. http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html (Accessed: 14 July 2024).

Contributors

Althage, Melanie, M.A., Research Assistant at the Chair for Digital History and at the Interdisciplinary Centre for Digitality and Digital Methods Campus Mitte (IC D2MCM) at Humboldt-Universität zu Berlin.

Anderson, Clifford, Dr., Director of the Divinity Library at Yale University.

Apel, Jochen, Dr., Director of Heidelberg University Library.

Bateman, John, Prof. Dr., Professor of English Applied Linguistics in the Department of Linguistics and Literature at the University of Bremen.

Burge, Caitlin, Dr., PostDoc with STEMMA, School of English, Media and Creative Arts, University of Galway, Ireland.

Burghardt, Manuel, Prof. Dr., Professor of Computational Humanities at the Institute of Computer Science at Leipzig University.

Ewerth, Ralph, Prof. Dr., Head of Visual Analytics Research Group at the TIB – Leibniz Information Centre for Science and Technology, Hannover.

Gius, Evelyn, Prof. Dr., Professor of Digital Philology and Modern German Literature at the Institute for Linguistics and Literary Studies at TU Darmstadt and first chair of the “DHd – Association for Digital Humanities in the German Speaking Areas” (“DHd-Verband”).

Gröpler, Johanna, M.A., Teaching and Learning Services, Deputy Head of Teaching Library and Reference Desk at FU Berlin.

Henny-Krahmer, Ulrike, Prof. Dr., junior academy professor for Digital Humanities at the University of Rostock.

Jannidis, Fotis, Prof. Dr., Professor for Digital Literary Studies at the University of Würzburg.

Kohle, Hubertus, Prof. Dr., Professor of Medieval and Modern Art History at the LMU Munich with a focus on digital art history.

Lasch, Alexander, Prof. Dr., Chair of German Linguistics and Language History at the TU Dresden.

Mattingly, William, Dr., Postdoctoral Fellow at the Smithsonian Institution Data Science Lab in collaboration with the United States Holocaust Memorial Museum (USHMM).

Müller-Budack, Eric, Dr. PostDoc in the Visual Analytics Research Group of the TIB – Leibniz Information Centre for Science and Technology, Hannover.

Mundorf, Margret, M.A., linguist, writing and communication trainer and lecturer at colleges and universities in Germany and Austria.

Nantke, Julia, Prof. Dr., Junior Professor of Modern German Literature with a focus on Digital Humanities for written artefacts at the University of Hamburg.

Nunn, Christopher A., Dr., PostDoc at the Chair for Historical Theology (Antiquity and Middle Ages) at Heidelberg University and head of the TheoLab.

Oorschot, Frederike van, PD Dr., lecturer in systematic theology at Heidelberg University, head of the “Religion, Law and Culture” department at the FEST Heidelberg and head of the TheoLab.

Peters, Janelle, PhD, Digital Humanities Chair of the North American Patristic Society (2022–2024), lecturer at Loyola Marymount University.

Piotrowski, Michael, Prof. Dr., Associate Professor of Digital Humanities in the Department of Language and Information Sciences in Faculty of Arts at the University of Lausanne (UNIL).

Raffety, Erin, PhD, cultural anthropologist, a Presbyterian pastor, and an ethnographic researcher at Princeton Theological Seminary.

Schroeder, Caroline T., Prof. Dr., Professor of Women’s and Gender Studies at the University of Oklahoma with a focus on Digital Humanities and Early Christianity, co-director of the Coptic Scriptorium.

Sprugnoli, Rachele, Prof. Dr., Assistant Professor in General and Historical Linguistics at the University of Parma.

Stökl Ben Ezra, Daniel, Prof. Dr., Research Professor at the Faculty of History and Philology of the École Pratique des Hautes Études – Paris-Sciences et Lettres Université (EPHE-PSL) in Paris.

von Stockhausen, Annette, Dr., Director of the academy project "The Late Antique Biblical Exegesis of Alexandria and Antioch" at the Research Centre for Primary Sources of the Ancient World of the Berlin-Brandenburg Academy of Sciences and Humanities.

Weiß, Christof, Prof. Dr., Professor for Computational Humanities at the University of Würzburg.

Wilder, Nicolaus, Dr., Research Associate at the Institute of Education at Kiel University.

Wilkens, Matthew, Prof. Dr., Associate Professor of Information Science at Cornell University.

Wunsch, Kevin, Research Assistant at the Centre for Digital Editions at the University and State Library at the Technical University Darmstadt.

Wuttke, Ulrike, Prof. Dr., Professor of Library Science – Strategies, Service Development and Scholarly Communication, Department of Information Sciences, University of Applied Sciences Potsdam, and board member of "DHd – Association for Digital Humanities in the German Speaking Areas" ("DHd-Verband").

Reviewer Wall of Fame

To express our gratitude to all those who have undertaken the arduous task of reviewing, we are pleased to present a list of those who have consented to be included in this wall of fame in alphabetical order.

Barker, Elton, Prof. Dr., The Open University (UK),
https://orcid.org/0000-0001-9517-1176.

Bernhart, Toni, Prof. Dr., University of Stuttgart,
https://orcid.org/0000-0002-7255-2504.

Du, Keli, Dr., Trier Center for Digital Humanities,
https://orcid.org/0000-0001-7800-0682.

Heßbrüggen-Walter, Stefan, Dr., University of Münster,
https://orcid.org/0000-0002-9383-026X.

Kampkaspar, Dario, University and State Library at the Technical University Darmstadt,
https://orcid.org/0000-0002-0118-0811.

Luth, Janine, Dr., Faculty of Modern Languages, Heidelberg University.

Ogihara, Mitsunori, Prof. Dr., Department of Computer Science, University of Miami, Coral Gables, FL, USA.

Pagel, Janis, Dr., University of Cologne,
https://orcid.org/0000-0003-4370-1483.

Petz, Cindarella, Dr., DH Lab, The Leibniz Institute of European History in Mainz,
https://orcid.org/0000-0002-6178-7332.

Revellio, Marie, Dr., University of Konstanz,
https://orcid.org/0000-0001-7264-5033.

Schröter née Vater, Christian, Dr., Academy of Sciences and Literature | Mainz – Digital Academy,
https://orcid.org/0000-0003-1367-8489.

Schumacher, Mareike, Jun.-Prof., University of Regensburg/University of Stuttgart,
https://orcid.org/0000-0002-7952-4194.

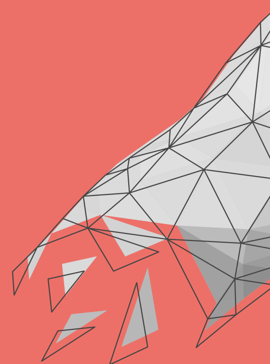
Seltmann, Melanie, Humboldt-Universität zu Berlin,
https://orcid.org/0000-0002-7588-4395.

Söring, Sibylle, Free University of Berlin,
https://orcid.org/0000-0002-1698-3289.

Spannagel, Christian, Prof. Dr., Heidelberg University of Education,
https://orcid.org/0000-0002-7299-4943.

Sutinen, Erkki, Prof. Dr., University of Turku,
https://orcid.org/0000-0002-1020-3325.

The Compendium of Computational Theology offers a first orientation at the interface of Theology and Digital Humanities. Volume 1 provides an overview of the objects, analytical tools and scholarly practices of the Digital Humanities. Volume 2 presents reflections and examples of how DH can be usefully applied in academic research and teaching in theological disciplines.



UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386