Virtuelle Forschungsumgebungen

Caroline T. Schroeder

https://orcid.org/0000-0001-9543-0692

Abstract Virtuelle Forschungsumgebungen in den theologischen Studien (und v. a. in den frühchristlichen Studien und den damit verbundenen Altertumswissenschaften) können eine wertvolle Infrastruktur für die Erstellung digitaler Editionen von Primärquellen und für andere Formen der digitalen und computergestützten Forschung bieten. Die Schaffung und Aufrechterhaltung dieser Umgebungen ist mit Herausforderungen verbunden. In diesem Beitrag werden die Vorteile der projektübergreifenden Zusammenarbeit sowie der gemeinsamen Nutzung und Wiederverwendung digitaler Ressourcen untersucht. Es werden auch einige Überlegungen zur Arbeit mit *unsauberen* oder *sauberen* digitalen Daten und zur Übernahme bestehender technischer Standards vorgestellt. In Bezug auf all diese Themen beinhaltet der Aufbau und die Nutzung von VREs die Entwicklung einer entsprechenden technischen Infrastruktur. Genauso wichtig wie die Technik sind jedoch die geisteswissenschaftlichen Fragen und die persönlichen Beziehungen, die einer erfolgreichen digitalen Initiative zugrunde liegen.*

Keywords Digital Humanities, virtuelle Forschungsumgebungen, Werkzeuge, Standards, Kollaboration, Open Access, Datenbereinigung, Frühchristliche Studien

1. Einführung

Virtuelle Forschungsumgebungen (orig. „Virtual Research Environments", VRE) in den theologischen Studien (und insbesondere in den frühchristlichen Studien und dem verwandten Bereich der Altertumswissenschaften) können eine wertvolle Infrastruktur für die Erstellung digitaler Editionen von Primärquellen und für andere Formen der digitalen und computergestützten Forschung bieten. Die Schaffung und Aufrechterhaltung dieser Umgebungen ist mit Herausforderungen verbunden. Zu den Schlüsselelementen erfolgreicher VREs gehören die projektübergreifende Kollaboration, die gemeinsame Nutzung und Wiederverwendung digitaler Ressourcen sowie die sorgfältige Abwägung, wie man mit *unsauberen* (orig. „messy") oder *sauberen* (orig. „clean") digitalen Daten arbeitet und ob man bestehende technische Standards übernehmen soll. In diesem Beitrag werde ich diese Aspekte der Arbeit in den Digital Humanities in unserem Forschungsfeld am Beispiel der Entstehungs-

* Dieses Kapitel wurde inkl. fremdsprachiger Zitate von der Redaktion aus dem Englischen übersetzt.

Caroline T. Schroeder: Virtuelle Forschungsumgebungen, in: Christopher A. Nunn und Frederike van Oorschot (Hrsg.): Kompendium Computational Theology 1. Forschungspraktiken in den Digital Humanities, Heidelberg: heiBOOKS, 2024, S. 411–426. DOI: https://doi.org/10.11588/heibooks.1459.c21924

geschichte der Plattform *Coptic Scriptorium* (CS) behandeln. Obwohl der Fokus dieses Beitrags auf dem CS liegt, werde ich auch andere VREs untersuchen und eine Analyse präsentieren, die über den Rahmen unserer individuellen Erfahrung hinausgeht.

Die virtuelle Forschungsumgebung *Coptic Scriptorium,* an deren Leitung ich beteiligt bin, entstand im Rahmen eines Sommerforschungsinstituts des *National Endowment for the Humanities* (NEH), das 2012 von der *Perseus Digital Library*[1] an der Tufts University veranstaltet wurde. Forschende aller Karrierestufen – von Doktorand*innen bis zu ordentlichen Professor*innen –, die in einer Vielzahl von Sprachen – Griechisch, Latein, Russisch, Koptisch – arbeiteten, bewarben sich und nahmen an einem dreiwöchigen Workshop teil, der von Monica Berti (Althistorikerin und Digital Humanist, jetzt an der Universität Leipzig), Gregory Crane (Tufts University, *Perseus*-Gründer) und Anke Lüdeling (Korpuslinguistik, HU Berlin) geleitet wurde. Zu diesem Zeitpunkt steckte das „digitale Koptisch" noch in den Kinderschuhen, und es gab nur wenige frei zugängliche VREs für frühchristliche Studien oder Altertumswissenschaften. Die *Perseus Digital Library,* unsere institutionelle Gastgeberin, war eine der bekanntesten (Crane 1998). *Trismegistos*[2] diente als *linked-data*-Struktur für Menschen, Orte und antike Texte (aufbauend auf und in Zusammenarbeit mit dem *Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens* [HGV] und der *Leuven Database of Ancient Books* [LDAB]) (Depauw & Gheldof 2014). *Papyri.info* hatte eine hochmoderne Umgebung für die kollaborative Textbearbeitung geschaffen, die von Crowd-Sourcing unter Papyrolog*innen profitierte.[3] Auch das *Tesserae*-Projekt an der University of Buffalo ist zu nennen, das gestartet wurde, um Forschungen zur Intertextualität in klassischen Quellen zu erleichtern. (Forstall et al. 2011; Okuda et al. 2022; vgl. auch den Beitrag von J. Nantke in diesem Band, S. 313). Es gab zwar weitere abonnementbasierte Forschungsumgebungen für Griechisch und Latein, aber nur wenige Open-Access- oder Open-Source-Umgebungen – die oben genannten sind einige der wichtigsten Projekte. Die Organisator*innen des NEH-Instituts hofften, dass die Teilnehmenden dort Anregungen fänden, um diese Lücken zu schließen.

In der Koptologie wurde der Unicode-Zeichensatz für das koptische Alphabet[4] im Jahr 2004 genehmigt und in den darauffolgenden Jahren um wichtige diakritische Zeichen erweitert wie z. B. im Jahr 2007 um verbindende Makron-Zeichen und das verbindende *Ni*, das am Ende von Zeilen in Manuskripten erscheint.[5] *Papyri.info* hatte kürzlich mit der Veröffentlichung einiger koptischer Papyri und Ostraka begonnen. Andere Institute und Einzelpersonen arbeiteten sowohl an koptischen als auch an

1 S. http://www.perseus.tufts.edu, zuletzt aufgerufen am 25.06.2024.
2 S. http://www.trismegistos.org, zuletzt aufgerufen am 25.06.2024.
3 S. http://papyri.info/ddbdp, zuletzt aufgerufen am 25.06.2024.
4 S. https://www.unicode.org/wg2/docs/n2824.pdf, zuletzt aufgerufen am 25.06.2024.
5 Zu den Revisionen im Jahr 2004 s. das Arbeitsblatt unter https://www.unicode.org/wg2/docs/n2744.pdf; zu 2007 s. http://unicode.org/wg2/docs/n3222 und https://www.unicode.org/L2/L2007/07118.htm [Protokoll des UTC 111/L 2 208 Joint Meeting]. Zur Standard Unicode-Schrift Antinoou (2012) s. http://www.evertype.com/fonts/coptic. Alle Adressen wurden zuletzt am 25.06.2024 aufgerufen.

syrischen Texten in Nicht-Unicode-Schriften und verbreiteten digitale Formen des Neuen Testaments und des christlichen Alten Testaments in diesen Sprachen (Schroeder 2019). Darüber hinaus war die jahrzehntelange Arbeit von Tito Orlandi am *Corpus dei Manuscritti Copti Letterari* (CMCL)[6] grundlegend (Orlandi 1997a; b; 2021). Dennoch standen nachhaltige digitale Editionen koptischer Literatur und nachhaltige digitale und computergestützte Forschung in der Koptologie erst am Anfang. Amir Zeldes, ein Linguist an der Humboldt-Universität (kein „Koptologe“) und ich (keine Linguistin), trafen uns am Tufts NEH-Institut, entdeckten unser gemeinsames Interesse an koptischer Literatur und Digital Humanities und begannen mit der Planung des Projekts. *Coptic Scriptorium*[7] startete 2013 mit einem ersten Pilotkorpus, Tools zur Verarbeitung natürlicher Sprache und einer einseitigen Website.[8] Inzwischen verfügen wir über eine Datenbank koptischer Literatur mit über 1,2 Millionen Wörtern (mit Anmerkungen zu Wortart, Syntax, Entitäten, Lemmata, Herkunftssprache, Manuskriptinformationen und mehr) sowie über mehrere Tools, darunter eine online NLP-Pipeline (Schroeder & Zeldes 2013–2023; 2016; 2020).

In diesem Beitrag werde ich drei Schlüsselthemen bei der Entwicklung von VREs ansprechen, die sowohl Herausforderungen als auch Chancen darstellten, als unser Projekt in den letzten zehn Jahren gewachsen ist: Spezialisierung und Zusammenarbeit bei der Wiederverwendung von Daten und Werkzeugen, unsaubere Daten und technische Standards. Der Aufbau und die Nutzung von VREs für die Digital Humanities-Forschung beinhaltet zwar die Entwicklung einer technischen Infrastruktur, aber ebenso wichtig ist es, geisteswissenschaftlichen Fragestellungen und kollaborativen persönlichen Beziehungen nachzugehen, die einer erfolgreichen digitalen Initiative zugrunde liegen.

2. Spezialisierung, Zusammenarbeit und Wiederverwendung

Digitale Forschungsumgebungen sind kostspielige Unternehmungen und oft ist das Publikum oder die Community der Nutzer*innen für solche Umgebungen überschaubar. In der Koptologie z. B. kennen sich die meisten von uns untereinander, egal ob wir in Nordamerika, Europa, Australien, Ägypten oder Japan arbeiten. Und es gibt wenig Raum für Überschneidungen in der Forschung – wenn wir bereits wissen, dass jemand an einer Edition bestimmter Handschriften oder Papyri arbeitet, geht der Rest von uns i. d. R. los, um an etwas anderem zu arbeiten. Dies hat sich auch in der

6 S. https://web.archive.org/web/19970624054528/http://rmcisadu.let.uniroma1.it/~cmcl, zuletzt aufgerufen am 25.06.2024.

7 S. https://copticscriptorium.org, zuletzt aufgerufen am 26.06.2024.

8 Obwohl wir keine Kopie der ursprünglichen Website mehr haben, ist die Version vom 9. Oktober 2014 in der *Internet Archive Wayback Machine* archiviert: https://web.archive.org/web/20141009102742/http://www.copticscriptorium.org, zuletzt aufgerufen am 25.06.2024.

digitalen Koptologie besonders bewährt. Das Ökosystem, das sich herausgebildet hat, besteht aus Spezialist*innen für bestimmte Bereiche. Und während die frühchristlichen Studien, die Klassische Philologie und die Bibelwissenschaften über weitreichendere Gemeinschaften verfügen, halten die Kosten für die Erstellung von VREs im digitalen Bereich von doppelter Arbeit ab. So spezialisieren sich die großen Open-Access-Projekte in der koptischen Text- und Sprachwissenschaft auf verschiedene Aspekte des Fachgebiets. Jede dieser Forschungsumgebungen hat sich als Reaktion auf die besonderen Forschungsbedürfnisse einer bestimmten Forschungsgemeinschaft entwickelt, und jede hat sowohl Einschränkungen als auch Vorteile.

Papyri.info veröffentlicht digitale Ausgaben von Ostraka und Papyri unter Verwendung der XML-Standards *(Extensible Markup Language),* die von der *Text Encoding Initiative* und der EpiDoc-Untergruppe der TEI entwickelt wurden (Elliott et al. 2006–2021).[9] Papyri und Ostraka sind in der Regel kürzer als literarische Texte und *Papyri.info* schafft eine digitale Forschungsumgebung, die mit den analogen Forschungsmethoden vergleichbar ist, die Papyrolog*innen traditionell anwenden (Editionen und Übersetzungen mit Notizen, Bildern, Apparaten usw.). Infolgedessen ist die Crowd-Sourcing-Digitalisierung von Papyri unter Papyrolog*innen möglich geworden. Sicherlich hat *Papyri.info* viel Zeit und Ressourcen in die Ausbildung und die Öffentlichkeitsarbeit investiert, was nicht unterschätzt werden kann; das Genre der Quellen und die digitalen Methoden tragen ebenfalls zu seinem Erfolg bei der Veröffentlichung einer enormen Anzahl von Dokumenten bei. Es gibt jedoch einige Merkmale, die in diesem Umfeld entweder nicht vorhanden sind oder mit denen es Probleme gibt (keine der folgenden Bemerkungen sollte als Kritik verstanden werden – es handelt sich um eine bemerkenswerte Leistung in Bezug auf Umfang und Methode. Die Beschreibung der Parameter der Plattform veranschaulicht, wie diese spezielle VRE spezifischen Forschungsfragen und -methoden dient). Die Plattform ermöglicht zwar die Suche nach einzelnen Wörtern und Wortreihen (einschließlich der Verwendung regulärer Ausdrücke) und bietet umfangreiche, durchsuchbare Metadaten, aber die Nutzer*innen, die ich auf Konferenzen kennengelernt habe, äußern gelegentlich Bedenken, dass die Ergebnisse einige Treffer auslassen oder dass sie nicht sicher sind, wie sie die Schnittstelle nutzen können, um so umfassende Suchergebnisse zu erzielen, wie sie es sich wünschen. Das Herunterladen von Ergebnissen für die rechnerische Arbeit ist für einfache Anwender*innen eine Herausforderung und die Wörter sind nicht mit einem Online-Wörterbuch wie in der *Perseus Digital Library* verknüpft. *Papyri.info* ist ein Kronjuwel der digitalen Altertumswissenschaften, weil es Funktionen enthält, denen sie sehr gut nachkommt. Keine Plattform kann jedoch alles für alle Nutzer*innen tun.

In ähnlicher Weise sehen wir eine Spezialisierung (und damit unterschiedliche Funktionen) in anderen frei zugänglichen VREs. In einem Projekt der Niedersächsi-

9 S. zu TEI http://www.tei-c.org; s. zu EpiDoc http://epidoc.stoa.org. Beide Adressen wurden zuletzt am 25.06.2024 aufgerufen.

schen Akademie der Wissenschaften zu Göttingen wird eine digitale Edition koptisch-sahidischer alttestamentlicher Handschriften[10] unter Verwendung der ursprünglich vom Institut für Neutestamentliche Textforschung geschaffenen Umgebung *Virtual Manuscript Room* erstellt, in der die Texte mit der Auszeichnungssprache TEI annotiert (Behlmer 2017). Das PAThs-Projekt in Rom hat einen *archäologischen Atlas der koptischen Literatur*[11] geschaffen, indem es eine Informationsstruktur für literarische Manuskriptdaten aufgebaut hat – wo Codices hergestellt und gefunden wurden, wo sie jetzt archiviert oder gelagert werden, wo sie veröffentlicht wurden, welche Werke auf jedem Codex erhalten sind usw. (Buzi 2017; Buzi et al. 2018). Der *Thesaurus Linguae Aegyptiae* (in Zusammenarbeit mit anderen) veröffentlichte ein ägyptisch-koptisches Lexikon im TEI-XML-Format, das vom CS in ein Online-Wörterbuch[12] umgewandelt wurde, und das Projekt *Database and Dictionary of Greek Loanwords in Coptic* steuerte anschließend seine griechische Lemmaliste und Definitionen bei (Feder et al. 2018; Burns et al. 2019).

Die Kollaboration mit anderen Projekten oder die Wiederverwendung ihrer Open-Source-Daten oder -Technologien ermöglicht es den Projekten, sich in ihren eigenen Forschungsbereichen hervorzutun, ohne das Rad in anderen neu erfinden zu müssen. Die meisten digitalen papyrologischen Projekte arbeiten mit *Papyri.info* zusammen, damit ihre Daten in die gemeinsame Datenbank einfließen können. Dies ermöglicht es Institutionen mit Papyrussammlungen, sich auf ihre spezifischen Objekte zu konzentrieren und gleichzeitig zu einer gemeinsamen Ressource beizutragen, von der eine breitere wissenschaftliche Gemeinschaft profitiert.

Das *Coptic Dictionary Online* (CDO) ist ein weiteres Beispiel für die Wiederverwendung und Kollaboration von Spezialist*innen. Es enthält Lexika aus zwei Projekten, zum einen das Wörterbuch und die Datenbank der griechischen Lehnwörter im Koptischen, zum anderen den *Thesaurus Linguae Aegyptiae.* Das CDO verlinkt jeden Wörterbucheintrag mit einzelnen Wörtern in den Korpora, die in der Datenbank des CS veröffentlicht sind; in ähnlicher Weise verlinkt die CS-Datenbank Wort für Wort zurück zum CDO. Darüber hinaus verweisen die Einträge für ägyptisch-koptische Wörter auf ein Online-PDF des umfangreichsten gedruckten koptischen Wörterbuchs (von Crum (1939), das von einem weiteren Partner, dem oben genannten alttestamentlichen Projekt aus Göttingen, bereitgestellt wird). Einträge zu griechischen Lehnwörtern verweisen auf das griechische Online-Wörterbuch *Perseus.* Das Team des CS entwickelte und pflegt die Online-Schnittstelle, die die Suche in der CDO und die Verknüpfung aller Ressourcen ermöglicht. Eine so umfassende, vernetzte und international weit verbreitete Ressource hätte von einer einzelnen Forschungseinheit allein nicht erstellt werden können.

10 S. https://www.uni-goettingen.de/en/digital+edition+des+koptischen+(sahidischen)+alten+Testaments/475974.html [zuletzt aufgerufen am 25.06.2024].

11 S. https://atlas.paths-erc.eu [zuletzt aufgerufen am 25.06.2024].

12 S. https://coptic-dictionary.org/about.cgi [zuletzt aufgerufen am 25.06.2024].

Solche Errungenschaften werden jedoch nicht ohne Herausforderungen erzielt. Im Koptischen z. B. sind sich die Koptolog*innen uneins darüber, was ein Wort in dieser Sprache ausmacht. Das mag geheimnisvoll klingen, aber diese Frage hat direkte Auswirkungen auf die Erstellung eines Online-Wörterbuchs. Koptisch ist eine agglutinierende Sprache, was bedeutet, dass verschiedene sprachliche Einheiten (z. B. ein Subjektpronomen und ein Verb) miteinander verbunden sind und zusammengeschrieben werden; außerdem werden koptische Manuskripte in *scriptua continua* geschrieben, d. h. ohne Leerzeichen zwischen den Wörtern oder verbundenen Wortgruppen. Die Segmentierung von Wörtern ist wichtig für die Suche und auch für die Erstellung lexikalischer Ressourcen, wie z. B. eines Wörterbuchs. Nehmen wir den Begriff für „Götzenanbeter", *refšmšeeidolon.* Sollen wir diesen Begriff als ein Wort mit einem lexikalischen Eintrag behandeln, da der gesamte Begriff sprachlich gesehen ein Substantiv ist, das einen bestimmten oder unbestimmten Artikel trägt und als ein Begriff Subjekt eines Verbs sein kann? Oder sollten wir es als drei Wörter behandeln, basierend auf den Morphemen, die den Begriff bilden *(ref-šmše-eidolon)*? Dabei bedeutet *šmše* „anbeten", *eidolon* ist „Götze", und *ref* ist die Vorsilbe, die anzeigt, dass ein Begriff ein Substantiv in der Form „die Person, die" das Folgende tut (die Person, die Götzen anbetet, oder „Götzenanbeter") ist. Die CS – mit ihrem Interesse an Linguistik, *Part-of-Speech*-Annotation und Syntax-Annotation – behandelt den Begriff als ein Wort (ein Substantiv) mit drei Morphemen. Die Forschungsinteressen der TLA bei der Erstellung ihres ägyptisch-koptischen Lexikons betreffen (teilweise) die Verfolgung der ägyptischen Sprache durch alle ihre Phasen. So behandelt es *ref-* als eigene lexikalische Einheit als Lemma und gibt ihm einen Eintrag im *Coptic Dictionary Online* („TLA lemma no. C3102"). Ein Klick auf den Link innerhalb dieses Eintrags, um Instanzen des „Wortes" *ref-* in der CS-Datenbank zu finden, führt jedoch *nicht* zu Treffern für alle Instanzen von *ref-* in unseren Korpora, da wir dieses Morphem als Präfix und nicht als Lemma oder Wort an sich behandeln; die Abfrage, die die CDO- und CS-Korpora-Datenbank miteinander verbindet, ist automatisiert, sodass die unterschiedlichen Datenmodelle in einigen wenigen Fällen (wie z. B. bei dem Morphem *ref-*) zu einer gewissen Unstimmigkeit führen.

Die Entscheidung über eine gemeinsame Definition dessen, was ein koptisches Wort oder Lemma ist, bevor das CDO gestartet wird, hätte diese Zusammenarbeit zum Erliegen gebracht. Stattdessen einigten sich die Projekte darauf, dass einige Unstimmigkeiten bei der Zuordnung unserer Daten ein kleiner Preis für den Gesamtnutzen der Verknüpfung des Wörterbuchs mit einer Online-Datenbank koptischer Textkorpora sind. Manchmal lassen sich diese Inkonsistenzen zumindest in eine Richtung auflösen; in der CS-Datenbank wird ein Wort wie *refšmšeeidolon* als drei Morpheme annotiert, wobei für jedes Morphem ein Link zum entsprechenden Eintrag im Online-Wörterbuch besteht. Man kann vielleicht nicht alle Treffer in der CS-Datenbank für Wörter, die mit *ref-* beginnen, mit einem Klick vom CDO-Eintrag aus erreichen, aber man kann sie mit einer leichten manuellen Änderung der Abfragesprache der Datenbank erhalten. Außerdem kann man den Wörterbucheintrag für *ref-* mit einem Klick

von der CS-Datenbank aus aufrufen. Die manuelle Zuordnung von Einträgen mildert einige andere Ungereimtheiten, aber eine solche Kodierung erfordert menschliche Arbeit, was angesichts der wettbewerbsorientierten und spärlichen Finanzierungsmöglichkeiten für viele geisteswissenschaftliche Projekte eine Herausforderung sein kann.

Die Anfänge von CS profitierten auch von der Nutzung früherer Arbeiten, einschließlich Open-Source-Technologie. Das Lexikon von Tito Orlandi (veröffentlicht im CMCL) ermöglichte es uns, Werkzeuge für die Verarbeitung natürlicher Sprache zu entwickeln, die den koptischen Text innerhalb des ersten Projektjahres in Wörter zerlegten und diese mit ihren Wortarten kennzeichneten. Dadurch konnten wir unsere ursprünglich angesetzte Arbeitszeit um ein Jahr verkürzen. Anstatt unsere eigene Datenbankinfrastruktur aufzubauen, adaptierten wir ein von Linguist*innen (darunter CS-Mitbegründer Zeldes) entwickeltes Open-Source-Tool (Zeldes et al. 2009; Krause & Zeldes 2014). Auch diese Wiederverwendung ermöglichte uns die Veröffentlichung eines durchsuchbaren Textkorpus innerhalb von Monaten anstatt Jahren. Andererseits kann die Suchoberfläche des Tools für Philolog*innen und Historiker*innen, die mit der Korpuslinguistik als Methode nicht vertraut sind, eine Herausforderung darstellen. Aus diesem Grund haben wir Online-Tutorials und Spickzettel bereitgestellt, um den Anwender*innen die Navigation im System zu erleichtern, und wir haben Entwicklungsressourcen in die Anpassung des Tools für das Koptische investiert. Auch wenn es nicht perfekt ist, überwiegen die Vorteile einer robusten, fast sofort einsatzbereiten Infrastruktur die Nachteile, v. a. mit Blick auf die Kosten für den Aufbau einer völlig neuen Datenbankinfrastruktur.

Zwangsläufig habe ich nicht alle VREs für Altertumswissenschaften oder frühchristliche Studien in diese Diskussion über Kollaboration und Wiederverwendung einbezogen. Dennoch veranschaulichen diese Beispiele einige der Herausforderungen, die sich aus der Spezialisierung, der disziplinären Vielfalt und den methodischen Unterschieden innerhalb der Disziplinen ergeben. Trotzdem können Open-Source- und Open-Access-VREs, die von Projekten verwaltet werden, die für Kollaborationen und Datenaustausch offen sind, weitaus solidere Forschungsmöglichkeiten bieten, als dies bei eher isolierten Projekten der Fall ist.

3. Unsaubere versus saubere Daten

Eine interdisziplinäre Debatte innerhalb der Digital Humanities, die sich direkt auf VREs in der Altertumswissenschaft und den frühen christlichen Studien auswirkt, ist die Frage, inwieweit wir unsere Textdaten bereinigen sollten. Die Philologie als Disziplin legt großen Wert auf Genauigkeit und Präzision bei Texteditionen wie auch bei Übersetzungen. Korpuslinguist*innen, Computerlinguist*innen und einige Digital Humanists haben eine höhere Toleranz für Unordnung.

Unter *unsauberen* geisteswissenschaftlichen Daten werden traditionell große Mengen an unstrukturierten und unbearbeiteten Texten verstanden (*big data,* Schöch 2013). Bis vor wenigen Jahren hatten Altertumswissenschaftler*innen nicht einmal Zugang zu „großen" antiken Textdaten in digitaler Form. Für Griechisch und Latein haben vor allem *Perseus* und *Open Philology,* aber auch andere Projekte zu einer umfassenden Digitalisierung beigetragen. Für Koptisch, Syrisch, Ge'ez und andere Sprachen bewegen wir uns langsam auf das zu, was wir als *medium data* bezeichnen könnten. Die digitale Altertumswissenschaft befindet sich in einem Spannungsfeld zwischen dem Wunsch nach größeren Korpora digitaler Daten, die wir durchsuchen oder analysieren können, einerseits und dem Vorrang hochpräziser, gründlich geprüfter Editionen andererseits. In einem Konferenzbeitrag aus dem Jahr 2013 über die Gründung und den langfristigen Fortbestand von *Papyri.info* nannte Roger S. Bagnall das *Peer-Review*-Verfahren als einen der Faktoren, die den Prozess der Veröffentlichung weiterer digitaler Editionen auf ihrer Plattform verlangsamen. Ein großer Teil von *Papyri.info* replizierte im digitalen Bereich – wenn auch in veränderter Form – die wissenschaftliche Form, die Papyrolog*innen gewohnt waren zu produzieren und zu benutzen – die Edition. Vor der Online-Veröffentlichung von Editionen wurde ein *Peer-Review*-Verfahren entwickelt, ähnlich wie bei gedruckten Editionen. Der Rückstand an Papyri oder Ostraka, die zur Veröffentlichung anstanden, wuchs bis zu dem Punkt, an dem der Projektvorstand beschloss, Editionen zu veröffentlichen, die noch nicht die letzte Runde der *Peer Review* durchlaufen hatten (Bagnall 2013). Bei digitalen Veröffentlichungen können wir natürlich schnell eine neue Version mit allen Korrekturen oder redaktionellen Änderungen herausgeben. Bei traditionellen gedruckten Ausgaben und Übersetzungen können Wissenschaftler*innen ein Jahrzehnt oder länger daran arbeiten, den Text mit detaillierten Anmerkungen oder Kommentaren zu versehen; außer bei sehr häufig gelesenen Werken ist das Erscheinen überarbeiteter Ausgaben oder neuer Ausgaben von anderen Wissenschaftler*innen kurz nach der vorherigen Veröffentlichung selten. Der Gründer von *Perseus*, Gregory Crane, kommentierte dieses Phänomen bereits in den 1980er Jahren in einem frühen Aufsatz über die Altertumswissenschaften und „Hypertext" (Crane 1987).

Im digitalen Zeitalter können *unsaubere* Daten eine Vielzahl von Dingen bedeuten, z. B. Ungenauigkeiten bei der optischen Zeichenerkennung (orig. „Optical Character Recognition", OCR) im Zuge der Digitalisierung von Druckausgaben, typografische Fehler bei der Transkription von alten Texten, typografische Fehler in Metadaten oder auch eine falsche Zuweisung von Quellen oder eine ungenaue Datierung. Bei Texten mit Anmerkungen zu sprachlichen Informationen wie Wortarten, Links zu anderen Ressourcen, Manuskriptinformationen usw. machen Fehler in den Anmerkungen die Daten ebenfalls „unsauber". Wissenschaftler*innen, die antike Texte redigieren, übersetzen und interpretieren, bringen oft zum Ausdruck, dass wir daran gewöhnt sind, mit sehr genauen Ausgaben zu arbeiten, was alle diese Aspekte betrifft – Genauigkeit des Textes, Informationen über das Werk, das den Text enthält, Übersetzung usw. In der Realität finden wir allerdings auch in Druckausgaben Fehler. Unsere Toleranz

für Fehler kann jedoch geringer sein als bei der Arbeit mit automatisierten digitalen Methoden. Genauigkeitsraten von 98–99 % für OCR gelten bspw. als recht hoch; bei einem Korpus von einer Million Wörtern bedeutet eine solche Rate, dass zehn- bis zwanzigtausend Zeichen betroffen sind – eine Zahl, an die sich Korpuslinguist*innen oder Informatiker*innen vielleicht gewöhnen, die aber viele Philolog*innen als beunruhigend empfinden könnten (zu OCR für historische Sprachen im Allgemeinen siehe Smith & Cordell 2018).

Einige Digital Humanists haben kürzlich Arbeiten veröffentlicht, in denen sie für mehr Toleranz gegenüber Unsauberkeit plädieren. Unsauberkeit kann *Ungenauigkeiten* in Daten oder Herausforderungen an hochstrukturierte, formale Systeme und Ideologien beinhalten, die einigen Computerarbeiten zugrunde liegen. Im letzteren Fall, so schreiben Losh et al. (2016), dient *Unsauberkeit* als theoretische Intervention in populäre Vorstellungen von digitalen Medien als ordentlich, sauber und hyperrational. In ähnlicher Weise argumentieren Katie Rawson und Trevor Muñoz, dass die Debatte über saubere vs. unsaubere Daten eine epistemologische ist: „Der Begriff ‚Bereinigung' impliziert, dass ein Datensatz zunächst ‚unsauber' ist. ‚Unsauber' suggeriert eine zugrundeliegende Ordnung: Es nimmt an, dass die Dinge bereits einen rechtmäßigen Platz haben, sich aber nicht an diesem befinden – wie Socken auf dem Schlafzimmerboden statt in der Kommode oder im Wäschekorb" (Rawson & Muñoz 2019). Aus dieser Sicht bedeutet das Bereinigen eines Datensatzes – insbesondere das Normalisieren oder Annotieren, um aus unstrukturierten *Daten* einen strukturierten Datens*atz* zu erstellen –, dass den Daten eine vorgefasste oder vorausgesetzte Ordnung oder ein Modell aufgezwungen wird. „Das Bereinigungsparadigma geht von einer zugrunde liegenden, ‚richtigen' Ordnung aus." Rawson & Muñoz (2019) plädieren dafür, die Vielfalt unsauberer Daten zu akzeptieren und zuzulassen, dass die Abfrage und Entdeckung *ungereinigter* Daten uns zu neuen Erkenntnissen über die Daten und die Gemeinschaften, die sie hervorgebracht haben, führt.

In der Philologie – und hier beziehe ich mich speziell auf die antike Literatur, insbesondere die Bibelwissenschaft, und nicht auf die Papyrologie – ist die Suche nach *sauberen* Textdaten mit der Suche nach dem *Urtext* verbunden. Mit *sauber* ist hier nicht der perfekt geschriebene oder genau kommentierte Text gemeint, sondern die früheste Version des Werks, die dem Original am nächsten kommt. Oft stimmt die sauberste kritische Ausgabe eines Werks mit keinem bekannten Manuskript zu 100 % überein. VREs und Methoden in der Handschriftenforschung verfolgen zwei unterschiedliche Ansätze für diese prädigitale Methodik. Werkzeuge und Projekte replizieren manchmal diesen traditionellen Prozess digital, indem sie Manuskriptzeugen transkribieren (oder VREs für die Transkription erstellen), die dann digital verglichen werden, um eine kritische Ausgabe zu erstellen (Behlmer 2017; Huskey 2019). Tools wie *Juxta Commons* und *CollateX*[13] ermöglichen es Forschern, parallele

13 S. https://collatex.net/about, zuletzt aufgerufen am 25.06.2024.

Zeugen desselben Textes während des digitalen Editionsprozesses zu markieren (Wheeler & Jensen 2014).

Einige Digital Humanists in der Klassischen Philologie haben auch untersucht, wie man den Druckapparat, den Philolog*innen zu sehen gewohnt sind, digital produzieren kann; als „Datenvisualisierung" ist der Apparat effizient und effektiv (Fischer 2019; Huskey 2022). Andere Projekte wie CS veröffentlichen digitale Ausgaben von Manuskripttranskriptionen (sowie frühere Druckausgaben) mit Metadaten, die Versionen desselben Werks miteinander verbinden, ohne jedoch einen Apparat oder eine kritische Ausgabe zu erstellen. Zumindest in dieser Hinsicht hat sich CS die *Unsauberkeit* zu eigen gemacht. Sicherlich ordnen wir den Text durch unsere linguistischen Anmerkungen, die ein Datenmodell verwenden, das zu einem großen Teil auf den grammatikalischen Kategorien und der Syntax in Bentley Laytons *Coptic Grammar* basiert – ein Werk, das selbst dafür kritisiert wird, dass es aggressiv neue linguistische Kategorien schafft und auferlegt (Layton 2011; Shisha-Halevy 2006; Feder 2017). Was jedoch die Editionen koptischer Literatur betrifft, so transkribieren wir bei der Veröffentlichung von Transkriptionen von Manuskripten den Originaltext (wie *unsauber* er auch sein mag) und erstellen einen normalisierten und lemmatisierten Text (die „saubereren" Textdaten) als Anmerkungen zum Original. So kann der Forschende nach einem erwarteten „sauber" geschriebenen Wort suchen und in unserer Datenbank alle Instanzen dieses Begriffs in seiner ursprünglichen Schreibweise sehen. In den Fällen, in denen wir sie veröffentlicht haben, können auch parallele handschriftliche Zeugnisse abgerufen werden. Wir bieten jedoch keine kritische Ausgabe oder einen Apparat.

4. Technische Standards

In den Digital Humanities haben technische Standards traditionell drei wichtige Funktionen. Standards legen den Grundstein dafür, wie Daten auszuzeichnen oder zu verarbeiten sind, damit nachfolgende Projekte das Rad nicht neu erfinden müssen. Auf diese Weise stellen sie eine gemeinsame Ressource für Geisteswissenschaftler*innen dar, die in verwandten Forschungsbereichen arbeiten. Meiner Meinung nach ist dies der wichtigste Aspekt digitaler Standards – eine Gemeinschaft kommt zusammen, um einen Fahrplan füreinander und für die Forschenden der Zukunft zu erstellen. Auch wenn nicht alle Aspekte der Standards für jedes einzelne Projekt in einem bestimmten Bereich geeignet sind, bieten sie einen Ansatzpunkt. Außerdem weisen sie andere Forscher*innen auf bekannte Probleme bei der Digitalisierung oder Berechnung in ihrem Fachgebiet hin. Das Datenmodell des PAThs-Projekts enthält zum Beispiel mehr als ein Feld für den*die Autor*in eines Werks – den*die „angegebene*n" Autor*in (wie im Manuskript oder Werk angegeben) und den*die „Urheber*in" (den*die nachweisbare*n historische*n Autor*in) (Buzi et al. 2018). Das Studium ihres Daten-

modells und ihrer Standards kann jedem Projekt helfen, das an Manuskripten und historischer Literatur arbeitet.

Theoretisch helfen Standards auch, die Konsistenz von Daten und Anmerkungen zu gewährleisten. Wenn z. B. geografische Orte in einem Datensatz auf die gleiche Weise annotiert sind, können Forschende nach einem Ort suchen und haben die begründete Erwartung, die meisten, wenn nicht sogar alle Instanzen dieses Ortes zu finden. Unterschiedliche Textdaten, die im Rahmen mehrerer Projekte nach demselben Standard annotiert wurden, können ebenfalls abgefragt und vergleichend analysiert werden. Ein solches Beispiel ist der *Universal Dependency Dataset* (UD), in dem Korpora aus über 100 Sprachen nach denselben linguistischen Standards annotiert wurden. Obwohl das Koptische lange Zeit als eine *unterversorgte* und vielleicht sogar obskure Sprache galt, bedeutet seine Präsenz im UD, dass Forscher*innen es neben modernen Sprachen wie Dänisch und Chinesisch untersucht haben, um Einblicke in die Sprache zu gewinnen (Zeldes & Abrams 2018; Pinter et al. 2019; Chen et al. 2022).

Schließlich sollte diese Konsistenz in der Theorie zu mehr Interoperabilität zwischen Projekten und Forschungsumgebungen führen. Digitale Editionen, die nach einem gemeinsamen Standard (wie TEI-XML) in einer VRE erstellt wurden, sollten in einer anderen VRE, die dieselben Standards verwendet, veröffentlicht oder bearbeitet werden können. *Papyri.info* ist ein solches Beispiel; es fasst Papyri und Ostraka, die von mehreren Projekten digitalisiert wurden, auf einer Plattform zusammen, was zum Teil durch die gemeinsame Nutzung der EpiDoc-Untermenge des TEI-XML-Standards möglich ist.

In der Praxis ist die Annotation jedoch ein Interpretationsprozess. Die Art und Weise, wie ein und derselbe Standard umgesetzt wird, kann variieren. Das CS, das Göttinger Projekt zum koptisch-sahdischen Alten Testament und das Projekt zu den Kanones von Apa Joannes dem Archimandriten haben sich alle auf die gemeinsame Nutzung von Daten geeinigt. Wir alle verwenden TEI-XML, um in unseren diplomatischen Transkriptionen Manuskriptinformationen zu vermerken. Allerdings verwenden wir einige der XML-Tags auf leicht unterschiedliche Art und Weise und wir haben auch unterschiedliche Auffassungen davon, was koptische Wörter in Phrasen bindet, die „verbundene Gruppen“ genannt werden. Daher haben wir Schriftkonverter erstellt, um eine echte Interoperabilität zu gewährleisten. Diese Unterschiede stellen keine kritischen oder unüberwindbaren Hindernisse für Kollaborationen dar, aber sie weisen auf das menschliche Element bei der gemeinsamen Nutzung von Daten hin. Darüber hinaus kann man in interdisziplinären Projekte feststellen, dass nicht alle Informationen, die im Rahmen ihres Projekts digitalisiert und kommentiert werden sollen, mit einem einzigen Satz von Standards erfasst werden können. Das CS gibt z. B. seine Daten in verschiedenen Formaten und nach unterschiedlichen Standards frei, da sich diese Standards in den einzelnen Fachbereichen für die jeweiligen disziplinären Bedürfnisse und Forschungsfragen entwickelt haben. Während TEI-XML ein robustes *Tagset* für digitale Editionen bietet, erfordert die Annotation von Wortarten und Syntax andere Arten von Auszeichnungen. Daher veröffentlicht unser Pro-

jekt unsere annotierten Korpora in verschiedenen Formaten. Jedes Dokument wird als „leichte" TEI-XML-Datei veröffentlicht, die Manuskriptinformationen und einige grundlegende linguistische Informationen (Herkunftssprache, Lemma, Wortart) enthält, als PAULA-XML-Dokumente mit vollständigen *Stand-Off*-Annotationen für alle Aspekte unseres Datenmodells (einschließlich kodikologischer und linguistischer Annotationen), als relationale Datenbankdateien, die vollständige Metadaten und Textannotationen enthalten, mit denen die ANNIS-Datenbank für die Abfrage unserer Korpora bestückt wird, und als SGML-Dokument mit allen Annotationen und Metadaten in einer Datei.[14] Wir generieren die Dateien in diesen verschiedenen Formaten aus einer Masterdatei. Außerdem veröffentlichen wir das oben erwähnte UD-Korpus, das eine Teilmenge unserer Korpora mit einem hohen Genauigkeitsgrad darstellt und gemäß den Syntaxstandards der UD-*Treebank* annotiert ist.

Kommunikation und Engagement für die Kollaboration innerhalb der Disziplinen und über disziplinäre Unterschiede hinweg sind ebenso wichtig wie technische Standards. Eine solche Kommunikation geht auch über den Bereich der Dokumentation hinaus. Die Dokumentation wird seit langem als Schlüsselelement für die Nachhaltigkeit und Nutzbarkeit von Projekten in den Digital Humanities genannt. Sie ist auch eine häufige Herausforderung, insbesondere für Projekte, die mit begrenzten Mitteln und/oder einem verkürzten Zeitrahmen für die Finanzierung laufen (Edmond & Morselli 2020). Die Standards eines Projekts sowie der Entscheidungsprozess oder die technischen Untersuchungen, die diesen Standards zugrunde liegen, können – und sollten – in Zeitschriftenartikeln, Projektblogs, Whitepapers und *„Read Me"*-Dateien dokumentiert werden. Es ist wichtig, transparent zu machen, wie eine VRE funktioniert, warum sie auf diese Weise funktioniert und wer zur Arbeit des Projekts beigetragen hat (Keralis et al. 2023). In kleinen Forschungsbereichen kultivieren erfolgreiche Projekte neben der Dokumentation von Standards eine menschliche Mentalität der Zusammenarbeit und der laufenden Kommunikation mit Nutzer*innen und Forschungspartner*innen.

5. Fazit

Viele Diskussionen über VREs oder andere *Werkzeuge* in den Digital Humanities drehen sich um Fragen der Nachhaltigkeit (vgl. den Beitrag von J. Apel in diesem Band, S. 438). Beim Aufbau eines Tools oder einer Plattform müssen die Projektteams den Arbeitsaufwand berücksichtigen, der für die Erstellung und den Support im Laufe der Zeit erforderlich ist, v. a. wenn sich Technologien und Standards ändern. VRE-Teams müssen sich Gedanken darüber machen, wie sie ausreichende Schulungen und Dokumentationen für die Benutzer*innen bereitstellen können. Nachhaltigkeit

14 S. https://github.com/CopticScriptorium/corpora, zuletzt aufgerufen am 25.06.2024.

ist nicht nur eine technische, sondern auch eine menschliche Frage. Die Entwicklung einer VRE, die flexibel genug ist, um über die anfängliche Startfinanzierung hinaus zu überleben (oder um Daten in Formaten zu produzieren, die überleben), erfordert sowohl technisches Fachwissen als auch persönliches Engagement für einen solchen Ansatz. Die Themen, die ich in diesem Kapitel angesprochen habe, sind in Gespräche über die Nachhaltigkeit der Digital Humanities eingebettet. Projekte, die VREs in den Digital Humanities einsetzen, können davon profitieren, wenn sie darüber nachdenken, wie sie bestehende Daten und Werkzeuge wiederverwenden können – und so den Lebenszyklus der Ergebnisse anderer Projekte verlängern und möglicherweise die finanziellen Kosten für die Entwicklungsarbeit in ihren eigenen Projekten reduzieren. Gespräche über technische Standards und *unsaubere* oder *saubere* Daten sind bei der Entwicklung von Plänen für die Beendigung eines Projekts unerlässlich. Die Planung der Zusammenarbeit von Anfang an kann Projekten dabei helfen, „das Rad nicht neu zu erfinden" und sie kann auch die Nutzung ihrer Daten oder Werkzeuge in einem größeren Rahmen und über einen längeren Zeitraum ermöglichen. Obwohl es sich bei einer VRE um eine technische Infrastruktur handelt, sind die Fragen und Methoden, die für den Aufbau und die Pflege eines solchen Instruments erforderlich sind, zutiefst menschlich.

Literaturverzeichnis

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 [zuletzt aufgerufen am 25.06.2024].

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. https://doi.org/10.1628/186870317X14876711440169 [zuletzt aufgerufen am 25.06.2024].

Burns, D. M., Feder, F., John, K., & Kupreyev, M. (2019). *Comprehensive Coptic Lexicon. Including Loanwords from Ancient Greek* [Datensatz]. https://doi.org/10.17169/REFUBIUM-2333 [zuletzt aufgerufen am 25.06.2024].

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. https://doi.org/10.1628/186870317X15100584934630 [zuletzt aufgerufen am 25.06.2024].

Dies., 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. https://doi.org/10.25592/uhhfdm.253 [zuletzt aufgerufen am 25.06.2024].

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 (Hrsg.), *The Co-Effect of Menzerath-Altmann Law and Heavy Constituent Shift in Natural Languages* (S. 11–24). Mouton: De Gruyter Mouton. https://doi.org/10.1515/9783110763560-002 [zuletzt aufgerufen am 25.06.2024].

Crane, G. (1987). From the Old to the New. Intergrating Hypertext into Traditional Scholarship. In *Proceedings of the ACM Conference on Hypertext* (S. 51–55). New York: Association on Computing Machinery. https://doi.org/10.1145/317426.317432 [zuletzt aufgerufen am 25.06.2024].

Ders. (1998). The Perseus Project and Beyond. How Building a Digital Library Challenges the Humanities and Technology, *D-Lib Magazine,* o. S. URL: http://www.dlib.org/dlib/january98/01crane.html [zuletzt aufgerufen am 25.06.2024].

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 (Hrsg.), *Theory and Practice of Digital Libraries. TPDL 2013. Selected Workshops.* Cham: Springer [= *Communications in Computer and Information Science,* 416]. https://doi.org/10.1007/978-3-319-08425-1_5 [zuletzt aufgerufen am 25.06.2024].

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). Rezension von Layton, Bentley. A Coptic Grammar, *Orientalistische Literaturzeitung,* 112(2), 108–12. https://doi.org/10.1515/olzg-2017-0035 [zuletzt aufgerufen am 25.06.2024].

Ders., 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* (S. 12–21). Santa Fe, New Mexico: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W18-4502 [zuletzt aufgerufen am 25.06.2024].

Fischer, F. (2019). Digital Classical Philology and the Critical Apparatus. In M. Berti (Hrsg.), *Digital Classical Philology. Ancienct Greek and Latin in the Digital Revolution* (S. 203–220). Berlin/Boston: De Gruyter Saur. https://doi.org/10.1515/9783110599572-012 [zuletzt aufgerufen am 25.06.2024].

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. https://doi.org/10.1093/llc/fqr029 [zuletzt aufgerufen am 25.06.2024].

Huskey, S. (2019). The Digital Latin Library. Cataloging and Publishing Critical Editions of Latin Texts. In M. Berti (Hrsg.), *Digital Classical Philology* (S. 19–34). Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783110599572-003 [zuletzt aufgerufen am 25.06.2024].

Ders. (2022). The Visual [Re]Presentation of Textual Data in Traditional and Digital Critical Editions, *Magazén,* 1. https://doi.org/10.30687/mag/2724-3923/2022/05/005 [zuletzt aufgerufen am 25.06.2024].

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 (Hrsg.), *Debates in the Digital Humanities* (o. S.). 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 [zuletzt aufgerufen am 25.06.2024].

Krause, Th., & Zeldes, A. (2014). ANNIS3. A New Architecture for Generic Corpus Query and Visualization, *Digital Scholarship in the Humanities,* 31(1), 118–139. https://doi.org/10.1093/llc/fqu057 [zuletzt abgerufen am 13.06.2024].

Layton, B. (2011). *A Coptic Grammar.* 3. Auflage. 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 (Hrsg.), *Debates in the Digital Humanities* (o. S.). Minneapolis: University of Minnesota Press. URL: https://dhdebates.gc.cuny.edu/read/untitled/section/cfe1b125-6917-4095-9d56-20487aa0b867#ch10 [zuletzt aufgerufen am 25.06.2024].

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 [zuletzt aufgerufen am 25.06.2024].

Orlandi, T. (2021). Reflections on the Development of Digital Humanities, *Digital Scholarship in the Humanities,* 36(2), 222–229. https://doi.org/10.1093/llc/fqaa048 [zuletzt aufgerufen am 25.06.2024].

Pinter, Y., Marone, M., & Eisenstein, J. (2019). Character Eyes. Seeing Language through Character-Level Taggers. In T. Linzen, G. Chrupala, Y. Belinkov, & D. Hupkes (Hrsg.), *Proceedings of the 2019 ACL Workshop BlackboxNLP. Analyzing and Interpreting Neural Networks for NLP* (S. 95–102). Florenz: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4811 [zuletzt aufgerufen am 25.06.2024].

Rawson, K., & Muñoz, T. (2019). Against Cleaning. In M. K. Gold & L. F. Klein (Hrsg.), *Debates in the Digital Humanities* (o. S.). URL: https://dhdebates.gc.cuny.edu/read/untitled-f2acf72c-a469-49d8-be35-67f9ac1e3a60/section/07154de9-4903-428e-9c61-7a92a6f22e51#ch23 [zuletzt aufgerufen am 25.06.2024].

Schöch, Ch. (2013) Big? Smart? Clean? Messy? Data in the Humanities, *Journal of Digital Humanities,* 2(3), o. S. URL: http://journalofdigitalhumanities.org/2-3/

big-smart-clean-messy-data-in-the-humanities [zuletzt aufgerufen am 25.06.2024].

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 (Hrsg.), *The Garb of Being. Embodiment and the Pursuit of Holiness in Late Ancient Christianity* (S. 318–345). New York: Fordham University Press.

Dies., & 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*]. https://doi.org/10.46298/jdmdh.5969 [zuletzt aufgerufen am 25.06.2024].

Dies. (2016). Raiders of the Lost Corpus, *Digital Humanities Quarterly*, 10(2), 1–38. URL: http://digitalhumanities.org/dhq/vol/10/2/000247/000247.html [zuletzt aufgerufen am 25.06.2024].

Schischa-Halevy, A. (2006). Rezension von Layton, Coptic Grammar. Second Edition, *Orientalia,* 75(1), 132–133. URL: https://arielshishahalevy.huji.ac.il/publications2006c [zuletzt aufgerufen am 25.06.2024].

Smith, D. A., & Cordell, R. (2018). *A Research Agenda for Historical and Multilingual Optical Character Recognition.* URL: http://hdl.handle.net/2047/D20297452 [zuletzt aufgerufen am 25.06.2024].

Wheeler, D., & Jensen, K. (2014). Juxta Commons [Poster], *Journal of Digital Humanities,* 3(1), o. S. URL: https://journalofdigitalhumanities.org/3-1/juxta-commons [zuletzt aufgerufen am 25.06.2024].

Zeldes, A., & Abrams, M. (2018). The Coptic Universal Dependency Treebank. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (S. 192–201). Brüssel: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6022 [zuletzt aufgerufen am 25.06.2024].

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/ [zuletzt aufgerufen am 25.06.2024].