



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 (= Input, Processing, Output): 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.

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-CRM⁴ 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. 92f.). 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 deteriorates. 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 occurrences 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-assisted 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 little 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 units, 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 addresses 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 as population 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.”

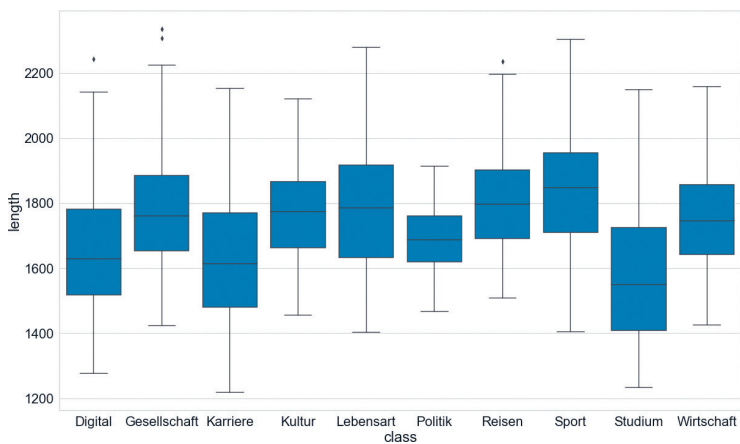


Fig. 1 Distribution of text length in the ten classes of newspaper articles.

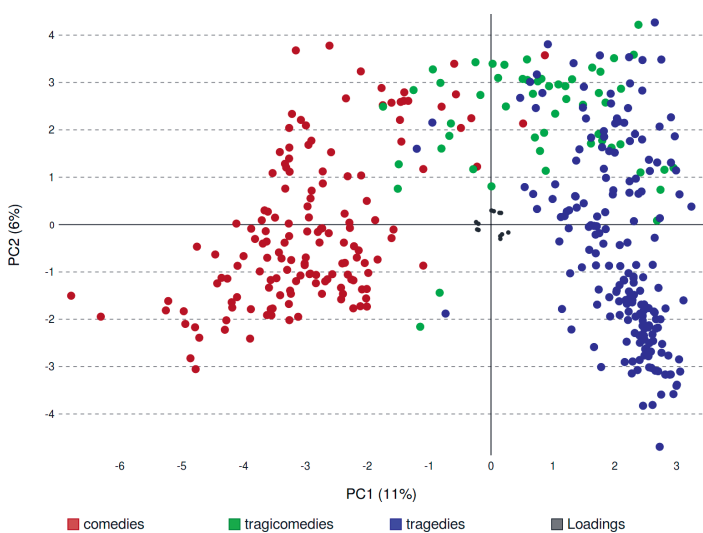


Fig. 2 Scatter plot after a Principal Component Analysis (a method of multivariate statistics) of French dramas and the likelihood of the occurrence of 60 themes within them.

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).