Methodenbeitrag: Digitale Editionen

Sandra Bläß
Marie Flüh [1]
Julia Nantke [1]
1. Universität Hamburg

Thema:	Textdigitalisierung und Edition	DOI:	10.48694/fortext.3747
Jahrgang:	1	Ausgabe:	3
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2022-07-04 auf fortext.net
Lizenz:	CC BY SA		open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

1. Definition

Digitale Editionen machen historische Dokumente für ein breites (wissenschaftliches) Publikum verfügbar und bilden damit die Basis für weitere Untersuchungen. Grundsätzlich können neben Textdokumenten auch kulturelle Artefakte in anderen medialen Formen wie audiovisuelle Medien oder bildnerische Objekte zum Gegenstand von Editionen werden. In der Literaturwissenschaft liegt der Fokus allerdings auf der Edition historischer Drucke und Handschriften. Sie werden in digitalen Editionen als digitale Faksimiles sowie als maschinenlesbare Transkripte repräsentiert und mit weiteren Informationen (vgl. Annotation) zu Überlieferung, relevanten Entitäten, Inhalten und/oder materiellen Besonderheiten angereichert. Die Bereitstellung erfolgt heutzutage zumeist über ein Online-Portal, wo die Dokumente im Open Access (vgl. Open Access) direkt rezipiert sowie (annotierte) Transkriptionen und Metadaten heruntergeladen werden können. Digitale Editionen sind nicht mit digitalen Repositorien zu verwechseln, die in der Regel nur einen maschinenlesbaren Text ohne Anbindung an die Überlieferung zur Verfügung stellen und nur wenig bis keine weiteren Informationen zum Text geben.

Editionen sind immer gleichzeitig Ergebnis wissenschaftlicher Arbeit und Basis für weitere Forschung. Um beide Perspektiven geht es im Folgenden.

2. Anwendungsbeispiel

Sie möchten Briefkorrespondenzen zwischen Künstler*innen um 1900 untersuchen und ein entsprechendes Korpus erstellen (Bläß 2024) – ein Ressourcentyp, der Ihnen Zugang zu wissenschaftlich valide aufbereiteten Texten gibt, sind digitale Editionen. Für das genannte Forschungsinteresse können Sie verschiedene Editionen kombiniert als Quellen nutzen, beispielsweise *Arthur Schnitzlers Briefwechsel mit Autorinnen und Autoren*, die Edition zu den Materialien um die Avantgarde-Zeitschrift *DER STURM* und das Portal *Dehmel digital*. Auf diesen Portalen können Sie beispielsweise über facettierte Suchen, durch Recherche im Register der erwähnten Personen, Orte, Werke und Organisationen oder über grafische Visualisierungen für Ihr Forschungsvorhaben relevante Texte auswählen. Weiterhin stehen dort Einzeldokumente für eine Detailansicht zur Verfügung.

Sollten Sie nicht nur mit einer bereits bestehenden digitalen Edition arbeiten, sondern selbst eine erstellen wollen, wenden Sie dabei, je nach Ausgangsmaterial und gewünschten Features (vgl. Feature), verschiedene Methoden an. Bei der Digitalisierung der analogen Dokumente kommen beispielsweise Varianten der Textdigitalisierung (Horstmann 2024a), der digitalen Manuskriptanalyse (Horstmann 2024b) sowie zur weiteren inhaltlichen Aufbereitung manuelle Annotation (Jacke 2024) oder Named Entity Recognition (Schumacher 2024) zum Einsatz.

3. Literaturwissenschaftliche Tradition

Die Editorik als Praxis des Herstellens zuverlässiger Textgrundlagen für die anschließende wissenschaftliche Rezeption hat eine lange Tradition, die bis in die disziplinären Anfänge der Geisteswissenschaften im 19. Jahrhundert zurückreicht. Editionen wurden zunächst von antiken und mittelalterlichen Texten angefertigt, die oftmals in vielen leicht voneinander abweichenden Fassungen überliefert sind. Ein Schwerpunkt der editorischen Arbeit lag deshalb auf dem Fassungsvergleich, um eine Annäherung an einen angenommenen ‚Urzustand'

des Textes zu erreichen. Mit zunehmender Relevanz moderner Texte für die historische und literaturwissenschaftliche Forschung verschob sich seit Beginn des 20. Jahrhunderts auch der editorische Fokus. Neben die differenzierte Aufschlüsselung der Überlieferung eines Textes rückte zunehmend die Textgenese sowie die dabei entstandenen Notizen, Entwürfe, und Reinschriften des/der Autors/Autorin in den Blick (Plachta 2020, 28–53). Auch hierbei ging und geht es bis heute weiterhin um eine textkritische Dokumentation des textuellen und materiell-medialen Befunds, der nach Möglichkeit von der Deutung der vorgefundenen Phänomene getrennt werden soll, auch wenn dies nicht immer vollständig möglich ist (Zeller 1971).

Zur Repräsentation von Überlieferungsgeschichten, Textgenesen und der Beziehung von Textfassungen und Varianten wurden von vornherein grafische Visualisierungen und diagrammatische Formen wie Baumdiagramme und Tabellen eingesetzt, wie sie heute vielfach in den DH üblich sind. Die Erschließungstiefe und der Umfang der Dokumentation und informativen Anreicherung einer Edition kann sehr unterschiedlich ausfallen. Im Laufe der Zeit haben sich drei archetypische Editionstypen herausgebildet, zwischen denen in der editorischen Praxis allerdings viele Mischformen bestehen (Plachta 1997, 11–26). Die (i) historisch-kritische Edition (HKA) richtet sich an ein wissenschaftliches Expert*innenpublikum. Sie erhebt den Anspruch, neben der Publikation eines zuverlässigen Texts auch den Zustand der Textgenese und/oder Überlieferung vollständig abzubilden, die Beziehungen zwischen den Dokumenten aufzuschlüsseln, deren jeweilige Authentizität und Position im Werk kritisch zu beleuchten sowie den aufgrund der Historizität der Texte bestehenden Verständnisschwierigkeiten durch gezielte Kommentierung zu begegnen. Ein Derivat der HKA stellt die (ii) Studienausgabe dar. Sie enthält ebenfalls zentrale textkritische und inhaltliche Erläuterungen, ist aber vorrangig auf den Gebrauch z. B. im Rahmen des Studiums ausgerichtet. Die einfachste Form bildet die (iii) Leseausgabe, die vor allem einen textkritisch geprüften, zuverlässigen Text enthält und auf weitere Informationen verzichtet. Im Idealfall können die drei genannten Editionstypen auseinander abgeleitet werden. Aus dieser Vorstellung hat sich im Rahmen der digital hergestellten Edition das heute etablierte Konzept des *Single Source Publishing* entwickelt, bei dem unterschiedliche Editionstypen ohne großen technischen Mehraufwand aus derselben Datenquelle erzeugt werden können (Sahle 2016, 32). In der im Internet publizierten Edition ist die Entscheidung darüber, wie viele und welche Informationen einem Dokument beigegeben werden können, nicht mehr durch materiell-mediale Erwägungen beschränkt, sondern hängt maßgeblich vom editorischen Konzept und den zur Verfügung stehenden Ressourcen ab.

4. Diskussion

Im Vergleich zu analogen Editionen bieten digitale Editionen einige Vorteile. Sie sind zeit- und ortsunabhängig nutzbare, zuverlässige Repräsentationen von Primärquellen, die allen Interessierten die Analyse wertvoller, empfindlicher und/oder sonst nicht ohne weiteres zugänglicher Dokumente ermöglichen. Digitale Editionen sind in der Regel kostenlos zugänglich und Sie müssen die Texte nicht selbst digitalisieren, um verschiedenste digitale Methoden darauf anwenden zu können. Zudem können digitale Editionen deutlich größere Bestände handhabbarer fassen als eine Printedition, sind nicht an eine einzelne Ordnungsstruktur gebunden, sondern können leichter verschiedene Perspektiven auf das Material anbieten, und sind nachträglich erweiter- und korrigierbar (Nutt-Kofoth 2016, 577–579). Diese Flexibilität impliziert aber auch potenzielle Schwierigkeiten bei der Zitation sich dynamisch wandelnder Inhalte, ein Problem, das jedoch nicht nur digitale Editionen betrifft: Allgemein sollte, so Föhr (2019), die Frage der Validität von und Kritik an Quellen aus dem Internet verstärkt in Studium und Forschung thematisiert werden. Bei digitalen Editionen gibt idealerweise ein Dokumentationsbereich des Portals Aufschluss über die Einhaltung wissenschaftlicher Standards, anhand derer Sie sich ein Bild machen können: Jede Edition sollte Editionsrichtlinien und eine Dokumentation enthalten, in denen Verarbeitungsschritte und Entscheidungen für oder gegen eine bestimmte Darstellung transparent vermittelt werden und erläutert wird, woher die Primärquellen stammen. Gute Editionsrichtlinien machen auch Lücken in der Erschließung transparent. Für die Zitation digitaler Editionen haben sich zwar noch keine einheitlichen Standards herausgebildet, allerdings beinhalten die meisten digitalen Editionen einen Zitiervorschlag, auf den Sie zurückgreifen können. Weitere Kriterien, anhand derer Sie digitale Editionen beurteilen können, finden Sie bei Sahle u. a. (2014).

Eine weitere Herausforderung besteht darin, dass technische Standards und Praktiken im stetigen Wandel sind: Anders als gedruckte Bücher veralten digitale Portale und können schlimmstenfalls nicht mehr nutzbar sein, wenn Nachnutzbarkeit im Projekt nicht ausreichend mitbedacht wurde. Diese Problematik wird im *Manifest für digitale Editionen* (Fritze 2022) u. a. darauf zurückgeführt, dass die zeitlich begrenzte Organisation und Finanzierung von Editionsvorhaben in wissenschaftlichen Projekten in Konflikt steht zur konzeptuellen Offenheit und Erweiterbarkeit digitaler Editionen.

Für die Nutzung digitaler Editionen spricht außerdem, dass sie meistens eine digitale Volltextsuche ermöglichen. Auf diese Weise gelangen Sie zügig zu den Dokumenten, die für Sie interessante, individuell wählbare Schlagworte enthalten. Viele digitale Editionen bieten zudem facettierte Suchen an. Auf diese Weise lassen sich große Textkorpora (vgl. Korpus) gezielt, also je nach individueller Forschungsfrage, innerhalb eines ausgewählten Zeitraums oder auf Grundlage weiterer flexibel kombinierbarer Parameter (vgl. Hyperparameter) durchsuchen (vgl.

Query). Beide Suchoptionen unterstützen Sie beim Auffinden von Dokumenten, bei denen sich ein Close Reading (vgl. Close Reading) für Sie lohnt (Baillot 2020, 392), und bei der gezielten Zusammenstellung eines Korpus, das für die Analyse und die Beantwortung Ihrer Untersuchungsfrage relevante Dokumente enthält. Letztere können Sie je nach Edition in unterschiedlichen Formaten wie XML oder PDF herunterladen. Einige digitale Editionen bieten zudem den Zugang über eine API an, über die der Gesamtbestand heruntergeladen werden kann. Nicht alle Downloadformate sind kompatibel mit weiteren Tools und Methoden der digitalen Textanalyse. Es kann also sein, dass Sie die heruntergeladenen Daten nach dem Download für weitere Analyseschritte anpassen müssen.

Editionen eröffnen stets einen neuen Zugang zum edierten Ausgangsmedium. Während dessen textuelle Ebene in digitale Volltexteditionen übertragen werden kann, verändert sich der materielle Zugang (Altenhöner u. a. 2014, 789), was durch Faksimiles und die Anreicherung mit weiteren Informationen zur Überlieferung und zum Zustand der Dokumente in Annotationen, Metadaten und/oder Kommentaren (vgl. Kommentar) in Teilen ausgeglichen werden kann. Eine digitale Edition kann in diesem Zusammenhang allerdings mit der Bereitstellung weiterer, alternativer Zugänge zum Material aufwarten. Mit den Dokumenten verlinkte Personennetzwerke (Dehmel digital), Wortwolken (vgl. Wordcloud)(Schlegel Edition), Visualisierung von Schritten der Textgenese (Faustedition) sowie Figurenauf- und Abtritten (Ödön von Horváth Digitale Edition) oder Landkarten schaffen Alternativen zum Einstieg über Suchfacetten, bieten alle Vorteile der Textvisualisierung (Horstmann und Stange 2024) und ermöglichen neue Perspektiven auf die Ausgangsdokumente.

Wichtiger Bestandteil aller Editionen sind Register: alphabetische Verzeichnisse wichtiger Einheiten. Während die Zusammensetzung (chronologische Auflistungen aller Personen, Orte, Werke und Körperschaften) und die Funktion der Register (Strukturierung des Materials) von digitalen und analogen Editionen sich nicht grundsätzlich unterscheiden und auch in der Handhabung ähnlich sind, werden Register digitaler Editionen in der Regel zusätzlich mit Verlinkungen auf Einträge in Normdatenbanken (vgl. Normdatenbank) angereichert: Hier sind weitere wissenschaftlich valide Informationen zu Personen, Orten oder Körperschaften gespeichert. Ist eine bestimmte Person also mit einem Eintrag in einer Normdatenbank verlinkt, gilt ihre Identität als eindeutig verifiziert. Darüber hinaus lassen sich digitale Editionen miteinander verlinken. Der Webdienst *correspSearch* ermöglicht die gleichzeitige Recherche in allen affiliierten Briefeditionen und gewährleistet eine editionsübergreifende Recherche, die dem Netzwerkcharakter vieler Korrespondenzen gerechter wird als der klassische Fokus auf die Briefe von ein und derselben Person (Weber 2013). Eine derartige Verknüpfung, die flexibel neu erscheinende Editionen aufnehmen und zugänglich machen kann, und generell die Möglichkeit von Anpassungen in der Edition, ist innerhalb von Printausgaben nur mit dem Zusatzaufwand neuer Auflagen zu leisten. Dies ist auch ein Schritt zur Lösung des Problems der Sichtbarkeit digitaler Editionen: Damit Sie sie nutzen können, müssen Sie sie erst einmal finden. Bislang sind digitale Editionen meist nicht in regulären Bibliothekskatalogen verzeichnet und unterschiedlich gut über Suchmaschinen auffindbar. Sie können jedoch im *Catalog of Digital Scholarly Editions* (Sahle u. a. 2020) recherchieren, der zahlreiche digitale Editionen und Editionsplattformen versammelt. Eine editionsübergreifende metadatenbasierte Recherche in Briefeditionen ermöglicht neben *correspSearch* auch der Kallilope Verbundkatalog. Archive, Bibliotheken, Museen und andere Einrichtungen (insgesamt Bestände aus über 950 Einrichtungen) hinterlegen hier u. a. Metadaten zu Korrespondenzen.

5. Technische Grundlagen

Für die Nutzung einer digitalen Edition benötigen Sie in der Regel keine technischen Vorkenntnisse, denn die Graphical User Interfaces (vgl. GUI) (GUIs) der meisten Editionen sind intuitiv nutzbar. Die edierten Dokumente werden auf der projekteigenen Homepage häufig synoptisch dargestellt (Faksimile neben Reintextfassung). Um eine effiziente facettierte Suche (vgl. Query) umzusetzen, ist manchmal eine bestimmte Syntax notwendig, die in der Regel aber erklärt wird.

Um Programmierschnittstellen (vgl. API) für den Download aller Dokumente einer Edition nutzen zu können, kann technisches Vorwissen notwendig sein. Grundkenntnisse zu TEI-XML (vgl. TEI) sind ebenfalls oft hilfreich, um aus einer digitalen Edition Ihr individuelles Textkorpus extrahieren und Daten aus verschiedenen Quellen aneinander angleichen zu können.

Sollten Sie nicht nur bereits bestehende digitale Editionen nutzen, sondern eine eigene erstellen wollen, müssen Sie sich dazu mit ganz unterschiedlichen konzeptionellen Fragen und digitalen Methoden auseinandersetzen, die im folgenden Teil kurz vorgestellt werden (Bläß u. a. 2022; Nantke, Bläß und Flüh 2022).

Die Textauswahl für digitale Texteditionen wird von mehreren Faktoren beeinflusst. Hierzu zählen beispielsweise urheberrechtliche Erwägungen, die Projektausrichtung und/oder förderpolitische Faktoren. Die fokussierte Textsorte sowie der Überlieferungszustand der Dokumente haben wiederum Einfluss auf die editorische Vorgehensweise sowie die Repräsentation auf dem Portal. Auch die Methoden unterscheiden sich je nach Zielen und (personellen, finanziellen und zeitlichen) Ressourcen des Projekts, in dem die digitale Edition entsteht. Es besteht zum einen die Möglichkeit, ausschließlich Scans plus Metadaten und Kommentar anzubieten; es gibt auch Editionen, die Volltexte ohne Faksimiles veröffentlichen. Im Weiteren beziehen wir uns vor allem auf

Volltexteditionen mit Faksimiles, die sich mittlerweile als Standard der digitalen Edition weitgehend etabliert haben. Auch in diesem Zusammenhang werden viele Briefeditionen hauptsächlich manuell erstellt. Das bedeutet: Das Dokument wird manuell transkribiert (vgl. Transkription). Dies kann erstens per Keying/Double Keying in Textverarbeitungsprogrammen, zweitens mit dafür vorgesehenen Tools wie Transcribo oder Transkribus oder direkt in einem Editor in TEI-XML geschehen. Die manuelle Annotation (Jacke 2024) der Transkripte beinhaltet zum einen Dokumenteneigenschaften und textkritische Merkmale wie Stiftwechsel, Durchstreichungen, die Textausrichtung oder, wenn mehrere Fassungen eines Textes vorliegen, variante Schreibweisen. Zum anderen beziehen sich die Annotationen auf inhaltliche Angaben. So werden z. B. bei Briefeditionen alle im Brief erwähnten Orte, Werke, Personen und Organisationen ausgezeichnet, sodass diese anschließend in Register übertragen werden können, die wiederum auf die Dokumente rückverlinken. In diesen Fällen sind die digitalen Editionen stark dem Druckparadigma verhaftet (Klug 2021) und unterscheiden sich bezüglich Layout, Aufbau und Nutzung kaum von analogen Editionen.

Alternativ zu dieser manuellen Vorgehensweise können Sie Verfahren des maschinellen Lernens (vgl. Machine Learning) wie Handwritten Text Recognition (vgl. HTR) (beispielsweise mittels Transkribus oder OCR4all) oder Named Entity Recognition (vgl. Named Entities) (z.B. mithilfe des Stanford Named Entity Recognizers) anwenden, um das Material editorisch zu erschließen. Darüber hinaus können Sie die spezifischen Möglichkeiten einer digitalen Umsetzung nutzen, indem Sie bspw. grafisch-visuelle Zugänge zum Bestand schaffen (bspw. über eine Netzwerkvisualisierung) und/oder Daten aus unterschiedlichen Quellen zusammenführen und miteinander verlinken. Sowohl die manuelle als auch die teilautomatisierte Vorgehensweise bedingt die Auseinandersetzung mit TEI-XML (siehe auch: https://tei-c.org40), dem Auszeichnungsstandard für digitale Editionen. XML-Dateien können Sie in verschiedenen Freeware-Texteditoren wie Notepad++ oder EMACS bearbeiten, für eine langfristigere und regelmäßige Nutzung kann sich jedoch auch Software mit zusätzlichen Features lohnen, beispielsweise der Oxygen-XML-Editor. Für diesen gibt es außerdem die an die Erstellung digitaler Editionen angepasste Arbeitsumgebung ediarum.

Je nach methodischer und inhaltlicher Schwerpunktsetzung unterscheiden sich Erschließungsbreite (der Gesamtumfang der in einer Edition repräsentierten Dokumente) und Erschließungstiefe (die Genauigkeit der Auszeichnung) bei manuellem und teilautomatischem Vorgehen: Editionen, bei denen ausschließlich manuell gearbeitet wird, sind in der Regel weniger umfangreich, da manuelles Transkribieren und Annotieren sehr zeitaufwändige Verfahren sind. Digitale Editionen dieser Art sind oft sehr genau ediert, da der Mensch bestimmte Schreibweisen oder Sinnzusammenhänge besser versteht als ein Computer. Dieses Vorgehen ist der editorische Standardfall. Editionen, bei denen Formen des maschinellen Lernens (vgl. Machine Learning) eingesetzt werden, können einen erheblich größeren Textbestand erfassen, da computationelle Methoden wie NER (Schumacher 2024) die Aufbereitung eines umfangreichen, potenziell grenzenlosen Datenbestandes ermöglichen, der mit manueller Herangehensweise nicht mehr (oder nur unter erheblichem Personal- und/oder Zeitaufwand) handhabbar wäre. Diese digitalen Editionen sind tendenziell weniger genau, da algorithmische Verfahren zum jetzigen Zeitpunkt oft Schriften und Entitäten nicht komplett fehlerfrei erkennen (Nantke, Bläß und Flüh 2022). Eine algorithmengestützte Editionspraxis zieht daher auch Kritik auf sich, da ein einseitiger Fokus auf innovative Technologien auf Kosten der inhaltlichen Qualität befürchtet wird (Rieger 2021).

Wenn Sie die unterschiedlichen Entitätentypen – Personen, Orte, Körperschaften und Werke – innerhalb Ihres für die digitale Edition vorgesehenen Textbestands (manuell oder automatisiert) digital annotiert haben, werden diese i. d. R. disambiguiert. Ein Tool, das die teilautomatisierte Beseitigung von sprachlichen Mehrdeutigkeiten (wenn z. B. „R.D.“, „Richard Dehmel“ und „Ri. Deh.“ immer als „Richard Dehmel“ im Register gelistet werden sollen) ermöglicht, ist bspw. OpenRefine. Über die „reconciliation“-Funktion können Sie dort auch die Vernetzung mit Normdaten vornehmen, da das Tool Schnittstellen zu unterschiedlichen Normdatenbanken wie Wikidata, der GND oder dem *Getty Thesaurus of Geographic Names* unterstützt.

Abgesehen von unterschiedlichen methodischen Schwerpunkten, die jedes Editionsprojekt selbst setzt, haben sich Standards etabliert, deren Einhaltung als gute wissenschaftliche Praxis gilt. Die 2016 eingeführten und interdisziplinär etablierten *FAIR Guiding Principles for scientific data management and stewardship* bilden die Grundlage für die Nachnutzung von Forschungsdaten und dienen auch digitalen Editionen als Bezugsrahmen (Stigler 2021): Sind die Daten auffindbar (*Findable*), zugänglich (*Accessible*), interoperabel (*Interoperable*, dazu zählt auch Bereitstellung der Dokumente in validem XML-Format und die Auszeichnung nach den Standards der *Text Encoding Initiative* (TEI)) und wiederverwendbar (*Reusable*), spricht das für eine digitale Edition mit hohen Qualitätsstandards. Es fehlen bislang jedoch einheitliche Standards für die Metadatenerfassung (Baillot 2020, 388).

Externe und weiterführende Links

- Arthur Schnitzlers Briefwechsel mit Autorinnen und Autoren: https://web.archive.org/save/https://schnitzler-briefe.acdh.oeaw.ac.at (Letzter Zugriff: 04.06.2024)

- correspSearch: https://web.archive.org/save/https://correspsearch.net/de/start.html (Letzter Zugriff: 04.06.2024)
- Dehmel Digital: https://web.archive.org/save/https://dehmel-digital.de (Letzter Zugriff: 04.06.2024)
- Dehmel Digital, Verlinkte Personennetzwerke: https://web.archive.org/save/https://dehmel-digital.de/network (Letzter Zugriff: 04.06.2024)
- DER STURM: https://web.archive.org/save/https://sturm-edition.de/ (Letzter Zugriff: 04.06.2024)
- ediarum: https://web.archive.org/save/https://www.ediarum.org (Letzter Zugriff: 04.06.2024)
- Faustedition: https://web.archive.org/save/http://www.faustedition.net/ (Letzter Zugriff: 04.06.2024)
- Kalliope Verbund: https://web.archive.org/save/https://kalliope-verbund.info/de/index.html (Letzter Zugriff: 04.06.2024)
- OpenRefine: https://web.archive.org/save/https://openrefine.org (Letzter Zugriff: 04.06.2024)
- Schlegel Edition: https://web.archive.org/save/https://august-wilhelm-schlegel.de/briefedigital/ (Letzter Zugriff: 04.06.2024)
- TEI: https://web.archive.org/save/https://tei-c.org (Letzter Zugriff: 04.06.2024)
- Ödön von Horváth Digitale Edition: https://web.archive.org/save/http://gams.uni-graz.at/context:ohad (Letzter Zugriff: 04.06.2024)

Bibliographie

Altenhöner, Reinhard, Tobias Beinert, Markus Brantl, Robert Luckfiel und Uwe Müller. 2014. Digitalisierung von kulturgut. In: *Praxishandbuch bibliotheksmanagement*, hg. von Rolf Griebel, Hildegard Schäffler, und Konstanze Söllner, 763–811. Berlin: De Gruyter Saur.

Baillot, Anne. 2020. Digitalisierung und ihre Einflüsse auf den Umgang mit alten wie neuen ‚Briefen' in deutscher wie internationaler Perspektive. In: *Handbuch Brief. Von der frühen Neuzeit bis zur Gegenwart*, hg. von Marie Isabell Matthews-Schlinzig, Jörg Schuster, Gesa Steinbrink, und Jochen Strobel, 387–395. Berlin/Boston: de Gruyter.

Berlin-Brandenburgische Akademie der Wissenschaften. 2021. ediarum. *ediarum*. https://www.ediarum.org/index.html (zugegriffen: 14. April 2022).

Bläß, Sandra. 2024. Methodenbeitrag: Korpusbildung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 2. Korpusbildung (12. Juni). doi: 10.48694/fortext.3708, https://fortext.net/routinen/methoden/korpusbildung.

Bläß, Sandra, Marie Flüh, David Maus und Julia Nantke. 2022. Quality Management for Machine Generated Data in Digital Scholarly Editions - Possibilities and Challenges. In: *Machine Learning and Data Mining for Digital Scholarly Editions*, hg. von Bernhard Geiger, Ulrike Henny-Krahmer, Fabian Kaßner, Marc Lemke, Gerlinde Schneider, und Martina Scholger. University of Rostock.

Bohnenkamp, Anne, Silke Henke und Fotis Jannidis. 2018. Faustedition. http://faustedition.net/ (zugegriffen: 4. Mai 2022).

Finkel, Jenny Rose, Trond Grenager und Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: *Proceedings of the 43nd annual meeting of the association for computational linguistics (ACL 2005)*, 363–370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf (zugegriffen: 12. April 2002).

Föhr, Pascal. 2019. *Historische Quellenkritik im Digitalen Zeitalter*. Glückstadt: Werner Hülsbusch.

Franz-Nabl-Institut, Karl-Franzens-Universität Graz. 2022. *Ödon von Horváth. Historisch-kritische Ausgabe – Digitale Edition*. http://gams.uni-graz.at/context:ohad (zugegriffen: 4. Mai 2022).

Fritze, Christiane. 2022. Manifest für digitale Editionen. *DHdBlog. Digital Humanities im deutschsprachigen Raum*. https://dhd-blog.org/?p=17563 (zugegriffen: 24. März 2022).

Horstmann, Jan. 2024b. Methodenbeitrag: Digitale Manuskriptanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3744, https://fortext.net/routinen/methoden/digitale-manuskriptanalyse.

———. 2024a. Methodenbeitrag: Möglichkeiten der Textdigitalisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3741, https://fortext.net/routinen/methoden/moeglichkeiten-der-textdigitalisierung.

Horstmann, Jan und Jan-Erik Stange. 2024. Methodenbeitrag: Textvisualisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3772, https://fortext.net/routinen/methoden/textvisualisierung.

Jacke, Janina. 2024. Methodenbeitrag: Manuelle Annotation. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3748, https://fortext.net/routinen/methoden/manuelle-annotation.

Klug, Helmut Werner. 2021. Editionstypen. Hg. von Helmut W. Klug, Selina Galka, und Elisabeth Steiner. *KONDE Weißbuch - Kompetenznetzwerk Digitale Edition*. hdl.handle.net/11471/562.50.76 (zugegriffen: 22. März 2022).

Kompetenzzentrum Trier Center for Digital Humanities. 2022. Transcribo. *Kompetenzzentrum Trier Center for Digital Humanities*. http://transcribo.org/de (zugegriffen: 12. April 2022).

Müller, Martin Anton. 2021. Arthur Schnitzler – Briefwechsel mit Autorinnen und Autoren. https://schnitzler-briefe.acdh.oeaw.ac.at/pages/index.html (zugegriffen: 12. April 2022).

Nantke, Julia. 2022. Dehmel digital. https://dehmel-digital.de/ (zugegriffen: 4. Mai 2022).

Nantke, Julia, Sandra Bläß und Marie Flüh. 2022. Literatur als Praxis. Neue Perspektiven auf Brief-Korrespondenzen durch digitale Verfahren. In: *Textpraxis. Sonderband: Digitale Verfahren der Literaturwissenschaft*, hg. von Frank Fischer und Jan Horstmann.

Nantke, Julia, Sandra Bläß, Marie Flüh und David Maus. 2022. Best of Both Worlds. Zur Kombination algorithmischer und manueller Verfahren bei der Erschließung großer Handschriftenkorpora. In: *DHd 2022*. Potsdam. doi: 10.5281/zenodo.6328113,.

Nutt-Kofoth, Rüdiger. 2016. Briefe herausgeben: digitale Plattformen für Editionswissenschaftler und die Grundfragen der Briefedition. In: *„Ei, dem alten Herrn zoll' ich Achtung gern". Festschrift für Joachim Veit zum 60. Geburtstag*, hg. von Kristina Richts und Peter Stadler, 575–586. München: Allitera.

Plachta, Bodo. 1997. *Editionswissenschaft*. Ditzingen: Reclam.

———. 2020. *Editionswissenschaft Handbuch zu Geschichte, Methode und Praxis der neugermanistischen Edition*. Stuttgart: Anton Hiersemann Verlag.

READ-COOP. 2021. Transkribus. *Transkribus. KI-gestützte handschriftenerkennung*. https://readcoop.eu/de/transkribus/ (zugegriffen: 12. April 2022).

Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner und Frank Puppe. 2019. OCR4all — An open-source tool providing a (semi-) automatic OCR workflow for historical printings. *Applied Sciences* 9, Nr. 22.

Rieger, Lisa. 2021. Digitale Edition. *KONDE Weißbuch - Kompetenznetzwerk Digitale Edition*. hdl.handle.net/11471/562.50.59 (zugegriffen: 22. März 2022).

Sahle, Patrick. 2016. What is a scholarly digital edition (SDE)? In: *Digital Scholarly Editing. Theory, Practice and Future Perspectives*, hg. von Matthew Driscoll und Elena Pierazzo, 19–39. Cambridge: Open Book Publishers. https://www.openbookpublishers.com/product/483/digital-scholarly-editing--theories-and-practices.

Sahle, Patrick, Georg Vogeler, Editorik und Institut für Dokumentologie und. 2014. Kriterienkatalog für die Besprechung digitaler Editionen. https://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/ (zugegriffen: 24. März 2022).

Sahle, Patrick, Georg Vogeler, Jana Klinger, Stephan Makowski und Nadine Sutor. 2020. A Catalog of Digital Scholarly Editions. https://www.digitale-edition.de/exist/apps/editions-browser/index.html (zugegriffen: 24. März 2022).

Schumacher, Mareike. 2024. Methodenbeitrag: Named Entity Recognition (NER). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 9. Named Entity Recognition (30. Oktober). doi: 10.48694/fortext.3765, https://fortext.net/routinen/methoden/named-entity-recognition-ner.

Stigler, Johannes. 2021. FAIR-Prinzipien. Hg. von Helmut W. Klug. *KONDE Weißbuch - Kompetenznetzwerk Digitale Edition*. https://www.digitale-edition.at/o:konde.7 (zugegriffen: 12. April 2022).

SyncRO Soft SRL. 2022. Oxygen XML editor. https://www.oxygenxml.com/ (zugegriffen: 12. April 2022).

Text Encoding Initiative. 2018. Text Encoding Initiative. *Text Encoding Initiative*. http://www.tei-c.org/index.xml (zugegriffen: 11. Juli 2018).

Trautmann, Marjam und Torsten Schrade. 2018. DER STURM. Digitale Quellenedition zur Geschichte der internationalen Avantgarde. *DER STURM. Digitale Quellenedition zur Geschichte der internationalen Avantgarde.* https://sturm-edition.de/ (zugegriffen: 12. April 2022).

Weber, Jutta. 2013. Briefnachlässe auf dem Wege zur elektronischen Publikation. Stationen neuer Beziehungen. In: *Brief-Edition im digitalen Zeitalter*, 34:25–34. editio Beihefte. Berlin u.a.: De Gruyter.

Zeller, Hans. 1971. Befund und Deutung. Interpretation und Dokumentation als Ziel und Methode der Edition. In: *Texte und varianten*, hg. von Gunter Martens und Hans Zeller, 45–90.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

API API steht für *Application Programming Interface* und bezeichnet eine Programmierschnittstelle, die Soft- und Hardwarekomponenten wie Anwendungen, Festplatten oder Benutzeroberflächen verbindet. Sie vereinheitlicht die Datenübergabe zwischen Programmteilen, etwa Modulen, und Programmen.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Close Reading Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen Annotation textueller Phänomene verbunden (vgl. auch Distant Reading als Gegenbegriff).

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

Double Keying Double-Keying ist eine Variante des Keying, bei der zwei Personen den Inhalt eines Dokumentes abtippen. Anschließend sucht ein Computerprogramm nach Differenzen zwischen den beiden Versionen. Gefundene Tippfehler werden dann von einer dritten Person korrigiert. So entstehen nahezu fehlerfreie Textdigitalisate.

Feature Unter Features können Einzelfunktionen eines Tools verstanden werden, die beispielsweise komplexe Funktionen wie die Visualisierung eines Textes als Wordcloud ermöglichen, oder auch kleinere Funktionseinheiten wie den Abgleich einzelner Spracheigenschaften (Properties) mit annotierten Beispieltexten darstellen.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

HTR HTR steht für *Handwritten Text Recognition* und ist eine Form der Mustererkennung, wie auch die OCR. HTR bezeichnet die automatische Erkennung von Handschriften und die Umformung dieser in einen elektronischen Text. Die Automatisierung beruht auf einem Machine-Learning-Verfahren.

Hyperparameter Hyperparameter beziehen sich auf externe, anpassbare Einstellungen, die genutzt werden um den Lernprozess zu kontrollieren und zu beeinflussen (zu modellinternen Parametern siehe Parameter). Sie sind unabhängig vom Datensatz und beziehen sich beispielsweise auf Einstellungen wie Anzahl der Iterationen, Größe der Datensätze oder Kontextfenster.

Keying In den Bibliotheks- und Textwissenschaften beschreibt Keying das manuelle Erfassen, also das Abtippen, eines Textes im Zuge seiner Digitalisierung (siehe auch Double-Keying).

Kommentar Textkommentare dienen meist der Erläuterung oder Interpretation literarischer Texte. Sie können entweder selbst Textform annehmen oder den Charakter von Anmerkungen haben. Treten sie in Form von Marginalien oder Glossen „in den Texten“ geschrieben auf, lassen sich auch Kommentare als Annotationen bezeichnen.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z.B. HTML,

zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

Normdatenbank Normdaten sind Datensätze, die Entitäten wie Personen, Werke der Literatur etc. eindeutig beschreiben und repräsentieren. Sie werden regelbasiert nach einer Normdatei angesetzt. Eine Normdatenbank ist demnach ein Datenbanksystem zur elektronischen Datenverwaltung von Normdaten.

OCR OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

Open Access Open Access bezeichnet den freien Zugang zu wissenschaftlicher Literatur und anderen Materialien im Internet.

Parameter Im Kontext von Machine-Learning-Modellen handelt es sich bei (Modell-)Parametern um modellinterne Konfigurationsvariablen, die anhand des Trainingssatzes bestimmt werden (zu modellexternen Parametern siehe Hyperparameter). Als Parameter werden einerseits Aspekte benannt, die den Lernprozess bestimmen und andererseits solche, die dabei erlernt werden. Die Werte der Parameter ergeben sich aus dem Datensatz selbst. Werte solcher Parameter können beispielsweise die Gewichtungen in neuronalen Netzwerken sein, also welche Aspekte im Trainingsprozess besonders einflussreich sind (z.B. können Wörter im direkten Umfeld eines Zielwortes als wichtiger bewertet werden also solche, die weit von diesem entfernt stehen) oder etwa wie die Gewichtung (also die Reihenfolge) der einzelnen Wörter innerhalb der Topics beim Topic Modeling.

PDF PDF steht für *Portable Document Format* . Es handelt sich um ein plattformunabhängiges Dateiformat, dessen Inhalt auf jedem Gerät und in jedem Programm originalgetreu wiedergegeben wird. PDF-Dateien können Bilddateien (z. B. Scans von Texten) oder computerlesbarer Text sein. Ein lesbares PDF ist entweder ein OCRter Scan oder ein am Computer erstellter Text.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Property Property steht für „Eigenschaft“, „Komponente“ oder „Attribut“. In der automatischen Annotation dienen konkrete Worteigenschaften wie Groß- und Kleinschreibung zur Klassifizierung von Wörtern oder Phrasen. Durch die Berücksichtigung solcher Eigenschaften in den Features eines Tools kann maschinelles Lernen bestimmter Phänomene umgesetzt werden. In der manuellen Annotation können als Properties auch Eigenschaften von Annotationen benannt werden.

Query *Query* bedeutet „Abfrage“ oder „Frage“ und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

Transkription Die Definition des Begriffs „Transkription” ist weit gefasst und stark abhängig vom wissenschaftlichen Bereich. Grundsätzlich bezieht sich die Transkription auf das Umschreiben, Übertragen oder Umformen einer Entität. In den Geisteswissenschaften kann sie grundsätzlich als Verschriftlichung von Medien wie Audio-, Videodateien aber auch Texten verstanden werden. In der Editionswissenschaft handelt es sich beispielsweise um die buchstabengenaue Abschrift oder Kopie eines Textes.

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

Wordcloud Eine *Wordcloud* , oder auch Schlagwortwolke, ist eine Form der Informationsvisualisierung, beispielsweise von Worthäufigkeiten in einem Text oder einer Textsammlung. Dabei werden unterschiedlich gewichtete Wörter, wie die häufigsten Wörter, i.d.R. größer oder auf andere Weise hervorgehoben dargestellt. Die horizontale/vertikale Ausrichtung und die Farbe der dargestellten Wörter hat meistens allerdings keinen semantischen Mehrwert.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.