Akademienunion

Berlin-Brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Offen für alle? Zugang gestalten in den Digital Humanities

Jahresworkshop 2026 der AG eHumanities der Union der deutschen Akademien der Wissenschaften

9.–10. November 2026
Berlin-Brandenburgische Akademie der Wissenschaften,
Einsteinsaal

Programm und Book of Abstracts

Inhaltsverzeichnis

Ankündigung und Call for Papers 1

Programm 3

Beiträge 5

- Session 1: Zugang, Offenheit & Barrierefreiheit 5
- Session 2: Sprache, Wörterbücher & KI 14
- Session 3: Digitalisierung & Quellenerschließung 23
- Session 4: Vernetzung, Knowledge Graphs & Suche 35
- Session 5: Forschungsdaten, Infrastrukturen & Zusammenarbeit 48

Berlin-Brandenburgische Akademie der Wissenschaften 2026, CC-BY SA 4.0
Gestaltung/Layout: Frederike Neuber



Foto von Patrick Robert Doyle auf Unsplash

Ankündigung und Call for Papers

Die Digital Humanities versprechen einen offenen Zugang zu Wissen, Daten und Forschungsinfrastrukturen. Zugang meint dabei nicht allein die freie Verfügbarkeit von Daten, sondern auch deren Auffindbarkeit, Verständlichkeit, technische Nutzbarkeit, rechtliche Nachnutzbarkeit und Anschlussfähigkeit. Doch wer hat tatsächlich Zugang zu digitalen geisteswissenschaftlichen Ressourcen – und wer bleibt ausgeschlossen? Welche technischen, institutionellen, sozialen und rechtlichen Faktoren fördern oder erschweren Zugänglichkeit? Und wie können digitale Forschungsangebote so gestaltet werden, dass sie unterschiedlichen Nutzer:innengruppen, Nutzungskontexten und technologischen Entwicklungen gerecht werden?

Der Jahresworkshop der AG eHumanities der Union der deutschen Akademien der Wissenschaften steht 2026 unter dem Titel „Offen für alle? Zugang gestalten in den Digital Humanities“. Mögliche Perspektiven umfassen soziale, technische, rechtliche und kulturelle Dimensionen von Zugänglichkeit: von Wissenschaftskommunikation, Open Access, Barrierefreiheit und Mehrsprachigkeit über Fragen der digitalen Teilhabe, Sichtbarkeit und Nachnutzbarkeit bis hin zur Gestaltung von Schnittstellen, Standards und Infrastrukturen, die den Zugang zu geisteswissenschaftlichem Wissen für Menschen ebenso wie für Maschinen ermöglichen.

Ein Schwerpunkt liegt auf Künstlicher Intelligenz und ihrer ambivalenten Rolle für die Zugänglichkeit digitaler geisteswissenschaftlicher Ressourcen: Verfahren der Text- und Handschriftenerkennung machen historische Quellen zugänglich und durchsuchbar, automatische Übersetzungen erleichtern den Umgang mit mehrsprachigen Inhalten, und KI-gestützte Bildbeschreibungen unterstützen inklusive Nutzungsszenarien. KI-Technologien können aber auch neue Barrieren schaffen, etwa durch ungleiche Sprachabdeckung, kostenpflichtige Dienste, technische Zugangsvoraussetzungen oder intransparente algorithmische Entscheidungsprozesse. Gleichzeitig gewinnt die Frage an Bedeutung, wie Wissen künftig digital erschlossen, vermittelt und rezipiert wird und welche Rolle Large Language Models dabei als neue Schnittstellen zu Wissen und kulturellem Erbe spielen. Dies gilt insbesondere für KI-Agenten als neue Nutzungsgruppe, für die der Zugang zu digitalen Angeboten eigens zu denken und zu gestalten ist.

Das Thema der diesjährigen Veranstaltung soll dazu anregen, darüber nachzudenken, wie Zugänglichkeit in den digitalen Geisteswissenschaften verstanden, gestaltet und nachhaltig gewährleistet werden kann und welche Herausforderungen dabei bestehen.

Wir begrüßen Einreichungen insbesondere zu folgenden Themen:

- Open Access, Open Science und offene Forschungsdaten
- Wissenschaftskommunikation und Public Humanities
- Community Building, Citizen Science und partizipative Ansätze
- Digitale Inklusion und Teilhabe unterschiedlicher Nutzer:innengruppen (z. B. in globalen und nicht-westlichen Forschungskontexten)
- User Interfaces und User Experience Design
- Barrierefreiheit digitaler Ressourcen und Anwendungen
- Multilingualität und mehrsprachige Wissensvermittlung
- Evaluation und Messung von Zugänglichkeit in digitalen Projekten
- KI-gestützte Lösungen zur Verbesserung der Zugänglichkeit (z. B. Bildbeschreibungen, Transkriptionen)
- Large Language Models und generative KI als Schnittstellen zu Wissen
- Maschinenlesbarkeit geisteswissenschaftlicher Daten
- Bias, Diskriminierung und Ausschlüsse durch algorithmische Systeme
- APIs, Linked Open Data und semantische Interoperabilität
- Nachhaltigkeit als Grundlage langfristiger Zugänglichkeit

Programm

V = Vortrag: 20 Min. Vortrag + 10 Min. Diskussion | LT = Lightning Talk: 10 Min. Talk + 5 Min. Diskussion

Montag, 9.11.2026

13:00–13:15 | Begrüßung (Thomas Elsaesser, Team AG eHumanities an der BBAW)

13:15–15:00 | Session 1: Zugang, Offenheit & Barrierefreiheit (Moderation: Frederike Neuber)

- Bärbel Kröger (AdW Göttingen): Eine offene Quelle, die niemand lesen kann: Ontologieentwicklung und Entitätendisambiguierung am Beispiel des Repertorium Germanicum (V)
- Eckhart Arnold (BAdW): Digitale Resilienz. Die Gretchenfrage des offenen Publizierens (LT)
- Matthäus Heil (Inscriptiones Graecae / BBAW): Schattenseiten von Open Access (LT)
- Johannes Ioannu (BBAW): Dieser Webauftritt ist nicht barrierefrei. Digitale Barrierefreiheit zwischen Anspruch und Praxis in den DH (LT)
- Nadine Quenouille (SAW Leipzig): One Document Does it all …: Nutzerzentriertes Design inklusive (LT)
- Stefan Dumont, Tobias Kraft (BBAW): Neue Wege in die Reisetagebücher der *edition humboldt digital* (LT)

15:00–15:30 | Kaffeepause

15:30–17:00 | Session 2: Sprache, Wörterbücher & KI (Moderation: Ralf Wolz)

- Daniel Deckers, Kajetan Dvoracek (Universität Hamburg / AdW Hamburg): Das Etymologika-Vorhaben: digitale Zugänglichmachung eines griechisch-byzantinischen Wörterbuchs (V)
- Amy Isard, Lena Schnee, Chris Biemann (Universität Hamburg): Eine neue Mittelniederdeutsche Grammatik (LT)
- Elena Lazarenko, Aleksandr Riaposov, Mariya Brykina, Beáta Wagner-Nagy (Universität Hamburg): Wie lassen sich KI-Tools für kleine nordeurasische Sprachen einsetzen? (LT)
- Anja Lobenstein-Reichmann, Lena Vosding, Jan Christian Schaffert (AdW Göttingen): EdiKILex. Edition, Künstliche Intelligenz und Lexikographie als Schlüssel zu vormodernen Textwelten (LT)
- Frank Wiegand (BBAW, DWDS): Offline für alle: Die DWDS-App (LT)

17:00–17:20 | Kaffeepause

17:20–18:50 | Session 3: Digitalisierung & Quellenerschließung (Moderation: tba)

- Petra Heinicker, Frederik Skidzun (BBAW; AdW Mainz): „Wenn Quellen flüstern. Zugänge zu beschädigten Quellen mittels Automated Text Recognition (ATR)“ (V)
- Moritz Eßer, Börger Kiss, Claes Neuefeind, Marcel Schaeben, Prianka Schmidt (Universität zu Köln; NRW Akademie): Digitalisierung zugänglich machen: Der Bildgebungs-Workflow für die hellenistischen Papyri der Bancroft Library (LT)
- Theresa Baumann, Maximilian Kristen (LMU München): Offen ist nicht gleich zugänglich. Das Corpus der barocken Deckenmalerei zwischen Retrodigitalisierung, Wissensvernetzung und KI (LT)
- Ute Tischer (SAW Leipzig): Digital anschlussfähig werden: Datenöffnung bei einer Printedition am Beispiel der Edition der Gottschedkorrespondenz (LT)
- Kevin Kuck, Kevin Wunsch (ULB Darmstadt): Einblick in die Automatisierungsworkflows in EuReD (LT)

ab 19:00 | Gemeinsames Abendessen (Erdinger, auf eigene Kosten)

Dienstag, 10.11.2026

09:00–10:30 | Session 4: Vernetzung, Knowledge Graphs & Suche (Moderation: Jörg Wettlaufer)

- Thomas Eckart, Uwe Kretschmer (SAW Leipzig): Offener Zugang trotz Komplexität: LLM-basierte Assistenzsysteme am Beispiel wissenschaftlicher Suchanwendungen (V)
- Peter Ackermann, Anna Martin (Steinheim-Institut, Universität Duisburg-Essen): Vernetzte Zugänglichkeit: Semantische Interoperabilität im Akademieprojekt „Steinerne Zeugen digital“ (LT)
- Hans Clausen (Freies Deutsches Hochstift): API, Filter-Interface und KI-Chat: Wie historia.editionhub.org einen Wissensgraphen befragt — Ein Werkstattbericht (LT)
- Linnaea Söhn, Martin Sievers, Torsten Schrade (AdW Mainz): AGATE meets KI: Einblicke in die experimentelle Öffnung des AGATE Knowledge Graphen mit LODEON (LT)
- Samuel Schepp, Sebastian Enns, Andreas Kuczera (TH Mittelhessen): Explorative Vernetzung verteilter Wissensquellen mit NAKAR (LT)

10:30–11:00 | Kaffeepause

11:00–12:45 | Session 5: Forschungsdaten, Infrastrukturen & Zusammenarbeit (Moderation: tba)

- Piroska Lendvai, Isabelle Mandrin (BAdW): Vom Zettel zur Quelle: Abrufsystem für ein lexikographisches Zettelarchiv mit der Anwendung „Zettel-Ranker“ (V)
- Hannah Busch (CCeH/Universität zu Köln): Lost in Translation? Über die Herausforderung der interdisziplinären Zusammenarbeit in Akademieprojekten (LT)
- Luis Sáenz (SAW Leipzig): Wer darf Infrastruktur bauen? KI-gestützte Entwicklung einer föderierten Bibliographie-Synchronisation zwischen zwei Akademievorhaben (LT)
- Marc Schulder, Thomas Hanke (Universität Hamburg): Anonym Gebärden? Privatsphäre im Hamburger Korpus der Deutschen Gebärdensprache (LT)
- Martin Prell (SAW Leipzig / Klassik Stiftung Weimar): Robustheitstests historischer Briefnetzwerke dank offener Forschungsdaten aus correspSearch, dem PROPYLÄEN-Projekt und dem Goethe- und Schiller-Archiv (LT)
- Julian Jarosch (AdW Mainz): Crowdsourcing-Plattformen als Projekt-Inkubatoren vor der Forschungsförderung (LT)

12:45–13:45 | Mittagsimbiss

13:45–14:45 | Keynote: Gitta Kutyniok (Professorin für Mathematische Grundlagen der Künstlichen Intelligenz an der Ludwig-Maximilians-Universität München, Akademiemitglied): Offen für alle – auch für KI-Agenten? Was vertrauenswürdige KI für den Zugang zu Wissen bedeutet (Moderation: Alexander Czmiel)

14:45–15:00 | Verabschiedung

Beiträge

Session 1: Zugang, Offenheit & Barrierefreiheit

Eine offene Quelle, die niemand lesen kann – Ontologieentwicklung und Entitätendisambiguierung am Beispiel des Repertorium Germanicum

Bärbel Kröger, Niedersächsische Akademie der Wissenschaften zu Göttingen

Vortrag

Der Beitrag gibt Einblick in das seit 1. Januar 2025 DFG-geförderte Verbundprojekt „Forschungsdateninfrastruktur Historische Quellen“ (HisQu). Beteiligt sind die Germania Sacra an der Niedersächsischen Akademie der Wissenschaften zu Göttingen, die Friedrich-Schiller-Universität Jena, das Deutsche Historische Institut in Rom und das Forschungszentrum Gotha der Universität Erfurt. HisQu entwickelt am Beispiel des Repertorium Germanicum (RG) Werkzeuge für die ontologiebasierte Quellenerschließung. Der Beitrag berichtet vor allem aus dem Göttinger Arbeitspaket, der Schwerpunkt liegt dabei auf der Modellierung, Erkennung, Typisierung und Disambiguierung historischer Entitäten.

Das vom Deutschen Historischen Institut in Rom herausgegebene RG erschließt die päpstlichen Register und Kameralbestände zur Geschichte des Heiligen Römischen Reiches (1378-1484). In rund 160.000 Regesten ist ein außerordentlich reiches Wissen über Rechtsakte, Personen, kirchliche Institutionen, Ämter, Verfahren, Orte und Beziehungen verdichtet. Als gedrucktes Werk entstanden, ist das RG zwar über eine Online-Datenbank frei verfügbar, aber dennoch ist die Regestensammlung äußerst schwer zugänglich. Die digitale Ausgabe ist keine semantisch tief erschlossene Edition. Die vorhandenen XML-Daten enthalten zwar einzelne Auszeichnungen; ein erheblicher Teil der Struktur bleibt aber in der Anordnung der Einträge, in editorischer Interpunktion, Formeln und hochgradig abgekürztem Latein implizit. Das RG ist deshalb nicht bloß ein Korpus schwieriger lateinischer Sätze. Seine Regesten bilden eine historisch und editorisch gewachsene Repräsentationssprache eigener Art. Punkte, Doppelpunkte, Positionswechsel und Abkürzungen markieren nicht nur sprachliche Grenzen, sondern können Rollen, Verfahrensstufen und Quellenreferenzen codieren. Wer diese Zeichen lediglich normalisiert, riskiert daher, Information zu beseitigen, die erschlossen werden soll (Camps/Vidal-Gorène/Vernet 2021). Der Zugang zum RG ist also nicht durch Lizenzen oder Schnittstellen beschränkt, sondern durch semantische Kompression – eine Barriere, die menschliche wie maschinelle Leser gleichermaßen trifft.

Um dieses Wissen explizit zu machen, wurde eine Architekturentscheidung getroffen, die im Projekt jüngst präzisiert wurde: Neben einer korpusnah erarbeiteten domänenspezifischen Ontologie entsteht mit der Factoid Core Ontology (FCO) eine generische Schicht, die Erwähnung und Referent trennt, konkurrierende und einander widersprechende Lesungen derselben Textstelle zulässt und Identifikationen samt Begründung modelliert (Pasin/Bradley 2015).

Die Domänenontologie ist genauer als fachwissenschaftliche Taxonomie beziehungsweise als kontrolliertes Vokabular zu verstehen, das von Historikerinnen und Historikern entwickelt und anschließend formalisiert wurde. Modelliert werden nicht allein die klassischen NER-Kategorien Person, Ort und Institution, sondern die für das Datenkorpus zentralen Rechtsakte und Benefizien wie Provision, Dispens, Absolution oder Ablass in ihrem Bezug zu kirchlichen Ämtern, beteiligten Rollen, zeitlichen

Angaben, Registerfundstellen und Überlieferungszusammenhängen. Definitionen und mögliche textuelle Manifestationen im RG werden in der Ontologie mit den Konzepten verbunden. Das Vorgehen setzte nicht ausschließlich bei vorab formulierten Competency Questions an (Noy/McGuinness 2001; Monfardini et al. 2023). Ausgangspunkt war ein Korpus, dessen begriffliche und relationale Vielfalt erst im Modellierungsprozess sichtbar wurde. Das Vokabular entstand daher korpusbasiert und iterativ aus den Regesten selbst sowie unter Rückgriff auf die vom Historischen Institut in Rom publizierten Editionsrichtlinien, Abkürzungsverzeichnisse und Register. Auch Ergebnisse veröffentlichter Forschung zum RG und das teilweise implizite Wissen langjähriger RG-Fachleute wurden bei der Entwicklung dieser domänenspezifischen Schicht der Ontologie mit einbezogen.

Dieser bottom-up orientierte Ansatz weist Berührungspunkte mit datengetriebenem Ontologielernen auf (Cimiano/Völker 2005), unterscheidet sich davon jedoch durch die fachwissenschaftlich kontrollierte Klassenbildung. Die ausdifferenzierte Hierarchie wirkte als Instrument der Quellenkritik: Sie zwang dazu, ähnliche Rechtsakte, institutionelle Typen und Akteursrollen zu unterscheiden und verstreutes editorisches Wissen zusammenzuführen. Die Domänen-Ontologie ist insofern ein epistemisches Werkzeug, mit dem das Korpus in seiner Eigenlogik sichtbar wird. Die Kernontologie und an sie anschließbare Inhaltsmodule aus der Domänen-Ontologie machen diese Tiefe infrastrukturell nutzbar. Der Kern verhindert eine bloß lokale Taxonomie ohne Anschlussfähigkeit; das Inhaltsmodul verhindert eine formal elegante, aber historisch zu grobe Modellierung. Die Ontologie wird nach Fertigstellung publiziert, insbesondere der domänenspezifische Teil dient zugleich als Glossar und Handreichung zum RG: ein Zugangsinstrument unabhängig von jeder Technologieentscheidung.

Ein anspruchsvoller Schritt ist die Überführung der Regesten in Instanzen der Ontologie, also die Ontologiepopulation. Dabei sind drei Aufgaben zu trennen, die in der Projektpraxis leicht vermischt werden: Named Entity Recognition bestimmt Textspannen und grobe Typen; semantische Extraktion rekonstruiert Rollen und Beziehungen; Entity Linking beziehungsweise Named Entity Disambiguation verknüpft eine Erwähnung mit einem bestimmten historischen Referenten. Ein System kann etwa einen Personennamen korrekt erkennen und dennoch die falsche Person identifizieren; ebenso kann es eine Kirche als Institution erkennen, aber Bittsteller, Empfänger und Referenzkirche eines Ablasses verwechseln. Für historische Forschung ist eine hohe Trefferzahl daher kein hinreichendes Qualitätsmaß. Entscheidend sind typ- und relationsspezifische Präzision, nachvollziehbare Kandidatenbildung, die Behandlung von Unsicherheit und die Möglichkeit, Entscheidungen auf den Regesttext und die verwendeten Normdaten zurückzuführen (vgl. Ehrmann et al. 2023).

Die Germania Sacra verfügt hierfür über eine ungewöhnlich dichte Wissensbasis. Dazu gehören Daten zu den Klöstern und Stiften des Alten Reiches, zu Bischöfen, Domherren, Diözesen und kirchlichen Ämtern, ferner Orts-Diözesen-Zuordnungen und Geodaten. Hinzu kommen bestehende Personen-, Orts- und Sachregister des RG, die editorischen Abkürzungsverzeichnisse sowie externe Referenzsysteme wie GND, Wikidata, GeoNames oder FactGrid. Diese Ressourcen enthalten bereits historische Identifizierungsarbeit. Sie nur als Trainingsmaterial für ein Sprachmodell zu behandeln, würde ihren epistemischen Status und ihre Provenienz verdecken. Im Workflow bilden sie deshalb zunächst regelbasierte Kandidatenräume. Schreibvarianten, institutionelle Typen, Diözesanzugehörigkeit, Amtszeiten, Datierungen und bekannte Beziehungen können mögliche Referenten eingrenzen oder ausschließen. Regeln ersetzen die historische Beurteilung nicht; sie machen aber explizit, weshalb ein Kandidat vorgeschlagen wurde, und erlauben eine reproduzierbare Fehleranalyse.

Das Vorgehen lässt sich an einem Ablassregest zeigen: „Aschersleve – eccl. domus fr. min. op. Halberstad. dioc.: indulg. ad instar eccl. s. Marie in Portiuncula 5 iul. 1401 L 94 199. (RG Online, Band 2, Nr. 645)“ Um diesen Eintrag zu erschließen, müssen mindestens der Ort Aschersleben, das

dortige Franziskanerkloster, die Diözese Halberstadt, der Rechtsakt Ablass, dessen Typ ad instar, die Referenzkirche Portiuncula, Datum und Registerfundstelle unterschieden werden. Ein reines Wörterbuchmatching bleibt dabei ebenso unzureichend wie ein generisches Sprachmodell: „Aschersleve“ bezeichnet im Regestkopf einen Ort, der Ablassempfänger ist aber eine konkrete Institution; die Kirche Portiuncula ist weder Ereignisort noch Empfänger, sondern Bezugspunkt des Rechtstyps. Erst die Verbindung von editorischer Segmentierung, Ontologie und domänenspezifischen Kandidatenregeln rekonstruiert diese Rollen.

Die verdichteten Formen des RG stellen Sprachmodelle vor eine besondere Herausforderung; ein eigener Test mit GLiNER (Costa/Kröger/Popp 2026; Zaratiana 2025) bestätigt dies. Ähnliche Beobachtungen liegen aus der digitalen Editorik vor (Camps u. a. 2021). Ein LLM-gestützter Workflow setzt voraus, dass zuvor expandiert wird – und hier liegt ein grundsätzliches Problem: Die Abkürzungsverzeichnisse sind über die Bände hinweg heterogen; dasselbe Kürzel kann bandabhängig Verschiedenes bedeuten. Die Auflösung ist damit kontextabhängig, also selbst eine Disambiguierungsaufgabe. Man benötigt für die Vorverarbeitung dieselbe Kompetenz, die das Verfahren erst herstellen soll. Hinzu kommt die Rekonstruktion der korrekten grammatikalischen Form aus der formelhaften Registersprache. Expansion ist folglich kein technischer Vorschritt, sondern ein editorischer Akt mit Interpretationsanteil. Belastbare Expansionsergebnisse erzielten wir bislang nur mit sehr leistungsfähigen proprietären Modellen.

Der hier nur knapp angerissene Forschungsstand zeigt, dass Evidenz material- und prozessschrittabhängig ist – was gegen jede generelle Methodenpräferenz spricht: Ehrmann u. a. (2023) zeigen, woran historische NER regelmäßig scheitert; Kovács et al. (2025) zeigen, dass ein speziell auf mittelalterlichem Latein vortrainiertes GLiNER-Modell nach dem Fine-Tuning praktisch keinen Vorteil gegenüber dem allgemeinen multilingualen Modell besitzt; Cui/Ströbel (2026) finden für lateinische Inschriften weiterhin feinjustierte BERT-Varianten vorn.

Für die Aufgabe der Entitätendisambiguierung im Projekt HisQu fällt die Wahl zunächst auf einen hybriden und modularen Workflow. Erstens wird die editorische Mikrostruktur segmentiert, ohne den Originaltext zu überschreiben. Zweitens werden sichere Entitäten über vorhandene Register und Normdaten erkannt; Kandidaten werden durch räumliche, diözesane, zeitliche und institutionelle Constraints gerankt. Drittens werden Abkürzungen nur kontextsensitiv und mit Provenienz expandiert, wobei Originalform, Kandidaten und gewählte Auflösung erhalten bleiben. Viertens kommen spezialisierte Modelle beziehungsweise Sprachmodelle dort zum Einsatz, wo Kontextinterpretation notwendig ist: zur Klassifikation seltener Rechtsakte, zur Rekonstruktion variabler Rollen oder zur Bewertung mehrerer plausibler Links. Fünftens werden Unsicherheit und konkurrierende Identifizierungen als Daten modelliert, statt durch einen einzigen scheinbar sicheren URI verdeckt zu werden. Schließlich dient fachwissenschaftliche Validierung nicht als nachträgliche Reparatur, sondern als Bestandteil der Methode und als Grundlage eines wachsenden Goldstandards. Sprachmodelle fungieren damit nicht als autonome Editoren, sondern als kontrollierte Komponenten in einem provenancebewussten Forschungsprozess.

Diese Architektur ist auch infrastrukturell relevant. Proprietäre Modelle liefern gegenwärtig teilweise bessere Resultate, schaffen aber Barrieren durch Kosten, Modellwechsel und intransparente Trainingsgrundlagen. HisQu untersucht deshalb offene Modelle und wissenschaftsnah betriebene Rechenangebote, unter anderem des Göttinger Rechenzentrums (GWDG). Offenheit erfordert zudem veröffentlichte Regeln, Versionen von Ontologie und Normdaten, Testkorpora, Fehlertypologien, Konfidenzen sowie maschinen- und menschenlesbare Entscheidungsprotokolle. Nur dann können andere Projekte Resultate prüfen, übertragen und korrigieren.

Der Beitrag argumentiert abschließend für einen relationalen Begriff digitalen Zugangs. Zwischen Quelle und Nutzer steht keine einzelne intelligente Schnittstelle, sondern eine Kette von Transformationen. Jede Stufe eröffnet Zugang und kann zugleich Ausschlüsse produzieren: Die Edition macht Registermaterial lesbar, ihre Abkürzungspraxis erschwert jedoch NLP-Verfahren; die Ontologie expliziert Fachwissen, erhöht aber die Einstiegsschwelle; Normdaten ermöglichen Verknüpfung, privilegieren jedoch gut dokumentierte Entitäten; Sprachmodelle erleichtern Interpretation, können Unsicherheit aber verdecken. Nachhaltiger Zugang entsteht deshalb durch die überprüfbare Kombination regelbasierter, ontologischer und fachwissenschaftlicher Verfahren mit KI. Für Menschen wie für KI-Agenten muss erkennbar bleiben, welche Zeichen im Regest standen, welche Transformation vorgenommen wurde, welche Wissensbasis eine Identifizierung stützt und wo historisches Urteil eingesetzt hat. Bei vormodernen Quellen bzw. Regestenwerken ist Transparenz nicht bloß Dokumentation, sondern Teil des Ergebnisses.

Literatur

- Camps, Jean-Baptiste; Vidal-Gorène, Chahan; Vernet, Marguerite (2021): Handling Heavily Abbreviated Manuscripts: HTR Engines vs Text Normalisation Approaches. In: ICDAR 2021 Workshops (LNCS 12917). Cham, S. 306–316. DOI: 10.1007/978-3-030-86159-9_21.
- Cimiano, Philipp; Völker, Johanna (2005): text2onto: A Framework for Ontology Learning and Data-Driven Change Discovery. In: NLDB 2005, 227–238.
- Costa, Luana Moraes; Kröger, Bärbel; Popp, Christian (2026): Schema-Driven Information Extraction from Medieval Latin Regesten: A Four-Way Evaluation of GLiNER2 for Ontology Population. SemDH 2026. https://ceur-ws.org/Vol-4212/short2.pdf.
- Cui, Wenhui; Ströbel, Phillip Benjamin (2026): „Across Generations: A Comparative Analysis of NER for Latin Inscriptions from Classical Machine Learning to LLMs“. In: Proceedings of LT4HALA 2026 @ LREC 2026. Palma, 112–124.
- Ehrmann, Maud; Hamdi, Ahmed; Pontes, Elvys Linhares; Romanello, Matteo; Doucet, Antoine (2023): Named Entity Recognition and Classification in Historical Documents: A Survey. ACM Computing Surveys 56/2, 1–47.
- Kovács, Tamás; Consolo, Giuseppe; Vogeler, Georg (2026): Twenty's Plenty: Semantic Scaffolding and Span Architecture for 19-Label NER in Medieval Latin Charters. In: Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities (NLP4DH 2026). ACL Anthology, S. 236-241. DOI: 10.18653/v1/2026.nlp4dh-1.22.
- Monfardini, Glaice K. Q.; Salamon, Jordana S.; Barcellos, Monalessa P. (2023): Use of Competency Questions in Ontology Engineering: A Survey. In: Almeida, J.P.A., Borbinha, J., Guizzardi, G., Link, S., Zdravkovic, J. (eds) Conceptual Modeling. ER 2023. Lecture Notes in Computer Science, vol 14320. Springer, Cham. https://doi.org/10.1007/978-3-031-47262-6_3.
- Noy, Natalya F.; McGuinness, Deborah L. (2001): Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report.
- Pasin, Michele; John Bradley (2015). „Factoid-Based Prosopography and Computer Ontologies: Towards an Integrated Approach.“ Digital Scholarship in the Humanities 30(1): 86–97. https://doi.org/10.1093/llc/fqt037.
- Shimizu, Cogan; Hammar, Karl; Hitzler, Pascal (2023): Modular Ontology Modeling. Semantic Web 14/3, 459–489.
- Zaratiana, Urchade; Pasternak, Gil; Boyd, Oliver; Hurn-Maloney, Ash; Lewis, Aaron (2025): GLiNER2: An Efficient Multi-Task Information Extraction System with Schema-Driven Interface. arXiv:2507.18546.

Digitale Resilienz. Die Gretchenfrage des offenen Publizierens

Eckhart Arnold, DH-Referat, Bayerische Akademie der Wissenschaften

Lightning Talk

Digitale Publikationen sind nach wie vor fragiler als Druckpublikationen. Nicht zuletzt, weil das Internet so gut funktioniert, existieren viele Digitale Publikationen nur in einer einzigen Fassung auf einem einzigen Server in der Welt. Langzeitbereitstellungskonzepte berücksichtigen häufig bestenfalls die zeitliche Dimension. Man vergleiche das mit Druck-Publikationen. Die finden sich in hunderten Bibliotheken der Welt. Brennt eine Bibliothek ab, so ist das Werk noch lange nicht verloren, während z.B. die Hälfte aller digitalen Editionen nach 13 Jahren nicht mehr im Netz zu finden ist. Man stelle sich einmal vor die Akademie-Ausgabe der Werke Kants wäre 13 Jahre nach ihrer Fertigstellung spurlos aus allen Bibliotheken der Welt verschwunden!

Offenbar gibt es bei digitalen Publikationen und Datenveröffentlichungen ein ungelöstes Resilienz-Problem. Wie könnte eine Lösung aussehen:

- Was sich bei Druck-Publikationen von selbst ergibt, sollte auch bei digitalen Publikationen angestrebt werden: Die Herstellung vielfacher geographischer und geopolitischer (!) Redundanz.
- Dieser Aufgabe müssten sich insbesondere die Bibliotheken als wissensbewahrende Institutionen annehmen, indem sie Spiegelserver für digitale Publikationen betreiben.
- Damit das funktioniert, müssen die Erzeuger ihre digitalen Werke so aufbereiten, dass sie immer auch in „statischer“ Form existieren können, d.h. einer Form, die ohne Server-Anwendung angezeigt werden kann, wobei auf ggf. Funktionen, nicht aber auf Inhalte und Darstellungsqualität verzichtet werden sollte. Hierfür eignet sich HTML/JS/CSS, mit größeren Einschränkungen auch PDF, nicht aber Quellen-Formate wie LaTeX oder TEI-XML.
- Bei Datenpublikationen bzw. Datenbanken ist das naturgemäß schwieriger. Hier könnte man ebenfalls über „sanft degradierte“ Fassungen nachdenken, die in Standard-Bibliothekssysteme eingespielt werden können. Oder entsprechende Datenprojekte sollte gleich als gemeinschaftliche Projekt mit weltweit verteilten Spiegelservern konzipiert werden.
- Nach wie vor eine gute Idee: Hybrid, d.h. auch Druckpublikationen. Aber bitte möglichst billig, denn das erhöht die Geo-Redundanz (vor allem im globalen Süden).Keinesfalls irgendwelche „renommierten“ Wissenschaftsverlage!

Literatur

- Kurzmeier, M., O'Sullivan, J., Pidd, M., Murphy, O. & Wessels, B., (2024) „Visualising the Catalogues of Digital Editions“, The Journal of Electronic Publishing 26(1). doi: https://doi.org/10.3998/jep.3569, URL: https://journals.publishing.umich.edu/jep/article/id/3569/.
- Harvard T.H. Chan School of Public Health. „As health data disappear from government websites, experts push back.“ Feb. 5, 2025. https://hsph.harvard.edu/news/as-health-data-disappear-from-government-websites-experts-push-back/
- E&E News / Politico. „All the climate info that disappeared under Trump. And how it’s being saved.“ E&E News, Jan. 7, 2026. https://www.eenews.net/articles/all-the-climate-info-that-disappeared-under-trump-and-how-its-being-saved/
- https://en.wikipedia.org/wiki/2025_United_States_government_online_resource_removals

Schattenseiten von Open Access

Matthäus Heil, Inscriptiones Graecae / Berlin-Brandenburgische Akademie der Wissenschaften

Lightning Talk

Open Access ist nicht gratis. Wie bei jeder Form des Publizierens fallen Kosten an. Mittlerweile werden sie zumeist auf die Autoren abgewälzt - ich habe es selbst erlebt. Für Wissenschaftsverlage ist dies ein risikoloses, lukratives Geschäftsmodell. Etliche Zuwendungsgeber übernehmen mittlerweile die Kosten, aber nur zu ihren je eigenen Bedingungen. Für Autoren stellt sich mitunter die Frage, was im Vergleich zu früher eigentlich besser geworden ist.

Es ist an der Zeit, die Kostenfrage offen und unvoreingenommen zu diskutieren - nicht um Open Access abzuschaffen, sondern damit Open Access wirklich funktionsfähig wird.

Literatur

- Open Access Geschäftsmodelle für Bücher (https://open-access.network/informieren/finanzierung/geschaeftsmodelle-fuer-buecher)
- Förderprogramm: „Open-Access-Publikationskosten“ | DFG (https://www.dfg.de/de/foerderung/foerdermoeglichkeiten/programme/infrastruktur/lis/lis-foerderangebote/open-access-publikationskosten)
- https://www.transcript-verlag.de/kosten-finanzierung

Dieser Webauftritt ist nicht barrierefrei. Digitale Barrierefreiheit zwischen Anspruch und Praxis in den DH

Johannes Ioannu, TELOTA, Berlin-Brandenburgische Akademie der Wissenschaften

Lightning Talk

Open Science heißt wissenschaftliche Forschungsdaten und Publikationen frei zugänglich und (nach-)nutzbar zu machen – ein Wissenschaftsverständnis, das auch für die Akademie leitend ist (BBAW 2022). Diese Ausrichtung bestärkt digitale Formate wie die digitale Edition, die neue Zugänge ermöglichen (Lavagnino 2009, 63). „Zugänglichkeit“ avanciert entsprechend zu einem „highly-cited term“ in den Digital Humanities (Cappellotto/Cioffi 2024, 2). Doch Offenheit allein garantiert noch keine Zugänglichkeit; und Zugänglichkeit selbst zeichnet sich durch eine „messy multiplicity“ an Lesarten aus (Martinez et al. 2019, 67). Ein Kernaspekt bleibt dabei oft wenig beachtet: die Frage der Barrierefreiheit digitaler Forschungsdaten und Publikationen. Zwar wird das Thema in den Digital Humanities vereinzelt schon länger thematisiert (vgl. Wymers 2005; Williams 2012; Del Turco 2011), eine systematische Bearbeitung, gerade im deutschsprachigen Raum, setzt aber vergleichsweise spät ein (Martinez et al. 2019, 55). Bis heute fehlt es an Strategien und Verfahren, um diesen Aspekt digitaler Zugänglichkeit nachhaltig zu sichern (Henzel 2023).

Dabei stehen Daten in keinem Widerspruch zu dieser Form von Zugänglichkeit – im Gegenteil (Anderson 2022, 2). Oft eröffnet erst die digitale Verfügbarkeit einer Ressource betroffenen Nutzer:innen überhaupt den Zugang zu den Inhalten; nicht umsonst zählte universeller Zugang für alle von Beginn an zu den Grundgedanken des Web (Berners-Lee 1998, 283). Viele Nutzer:innen mit Beeinträchtigungen sind auf gut strukturierte, sauber ausgezeichnete Daten angewiesen und im Umgang mit ihnen entsprechend versiert (Anderson 2022, 2). Die Grundlage barrierefreier digitaler Ressourcen liegt somit dort, wo gute wissenschaftliche Praxis im Digitalen ohnehin ansetzt: bei sorgfältig und konsistent annotierten Daten.

Schwieriger wird es bei der Frage, wie der Zugang zu diesen Daten gestaltet wird. Fehlt es an dieser Stelle an der notwendigen Sensibilisierung für eine inklusive Gestaltung, entstehen Barrieren. Das gilt in besonderem Maße für die deutsche Wissenschaftslandschaft, in der Menschen mit Behinderungen unterrepräsentiert sind und eine gleichberechtigte Teilhabe „noch immer nicht selbstverständlich ist“ (Beer et al. 2025, 674).

Wollen wir in den Digital Humanities die erarbeiteten Inhalte wirklich zugänglich machen, müssen wir uns für Barrierefreiheit sensibilisieren und sie praktisch umsetzen. Gute Ausgangspunkte sind Weboberflächen. Hier bieten etablierte Standards wie die Web Content Accessibility Guidelines (WCAG) (Campbell et al. 2023) mit den darin formulierten POUR-Prinzipien eine erste Orientierung, um zu verstehen, wie und wo im eigenen Projekt Barrieren entstehen können. Zudem lassen sich diese Kriterien systematisch überprüfen. Ob eine Anwendung tatsächlich barrierefrei nutzbar ist, zeigt sich jedoch erst in Tests unter Einbezug von Expert:innen und betroffenen Nutzenden. Wie digitale Ressourcen in den DH selbst ist auch Barrierefreiheit kein Produkt, sondern ein Prozess, der umso mühsamer wird, je später er einsetzt. Ansätze wie Universal Design bzw. Inclusion by Design betonen daher die Notwendigkeit, Fragen der Barrierefreiheit von Anfang an im Projekt mitzubinden – wovon am Ende alle Nutzer:innen profitieren.

Der Lightning-Talk versteht sich als Diskussionsimpuls für diesen Prozess. Auf Basis der von Kathryn Wymer bereits 2005 formulierten drei Gründe, warum Barrierefreiheit für eine digital arbeitende Geisteswissenschaft von Interesse sein sollte, nimmt der Vortrag eine Neubewertung aus heutiger Sicht vor. Webauftritte digitaler Ressourcen der Berlin-Brandenburgischen Akademie dienen dabei als Anschauungsmaterial, ergänzt um Erfahrungen und Ergebnisse aus der Arbeit zur digitalen Barrierefreiheit bei TELOTA. Den Abschluss bildet ein Überblick über bestehende Initiativen und Anknüpfungspunkte im Fach sowie ein Minimalvorschlag für einen ersten Ansatz, um Barrierefreiheit in DH-Projekten stärker in den Blick zu nehmen. Barrierefreiheit entsteht dabei nicht durch exklusive Entscheidungen, sondern nur im kontinuierlichen und intensiven Austausch aller Beteiligten. Denn offen für alle ist erst, was auch barrierefrei ist.

Literatur

- Anderson, Theresa, Randy D. Colón, Abigail Goben, und Sebastian Karcher. 2022. „Curating for Accessibility“. International Journal of Digital Curation 17: 10. https://doi.org/10.2218/ijdc.v17i1.837.
- Beer, Anna, Simone Franz, Katrin Henzel, und Jan Klaes. 2025. „Inklusion und Forschungsdaten: GO UNITE! AG Inklusion im Forschungsdatenmanagement“. In Praxishandbuch Inklusion in Bibliotheken. Barrierefreier Zugang zu Information, Bildung und Kultur, herausgegeben von Christiane Felsmann, Belinda Jopp, und Anne Sieberns. De Gruyter Saur. https://doi.org/10.1515/9783111206943-086.
- Berners-Lee, Tim. 1998. „Realizing the potential of the Web“. In Web-Weaving. Intranets, extranets and strategic alliances, herausgegeben von Peter Lloyd und Paula Boyle. Oxford.
- Campbell, Alastair, Chucks Adams, Rachael Bradley Montgomery, Michael Cooper, und Andrew Kirkpatrick. 2023. „Web Content Accessibility Guidelines (WCAG) 2.2“. Informationsportal. W3C. https://www.w3.org/TR/WCAG22/.
- Cappellotto, Anna, und Raffaele Cioffi. 2024. „Towards an Inclusive and Accessible Digital Scholarly Editing: A Critical Assessment“. Journal of the Digital Humanities Association of Southern Africa (DHASA), Digital Humanities for Inclusion, Bd. 5 (1). https://doi.org/10.55492/dhasa.v5i1.5012.
- BBAW. 2022. „Das Leitbild Open Science der Berlin-Brandenburgischen Akademie der Wissenschaften“. Workingpaper. https://nbn-resolving.org/urn:nbn:de:kobv:b4-opus4-37530
- Henzel, Katrin. 2023. „Vermittlung Auf Augenhöhe – Digitale Editionen Inklusiv Gestaltet“. Editio 36 (1): 72–88. https://doi.org/10.1515/editio-2022-0003.
- Lavagnino, John. 2009. „Access“. Literary and Linguistic Computing 24 (1): 63–76. https://doi.org/10.1093/llc/fqn038.

- Martinez, Merisa, Wout Dillen, Elli Bleeker, Anna-Maria Sichani, und Aodhán Kelly. 2019. „Refining our conceptions of ‘access’ in digital scholarly editing: Reflections on a qualitative survey on inclusive design and dissemination“. Variants. The Journal of the European Society for Textual Scholarship 14: 41–74. https://doi.org/10.4000/variants.1070.
- Rosselli Del Turco, Roberto. 2011. „After the Editing is Done. Designing a Graphic User Interface for Digital Editions“. Digital Medievalist 7. https://doi.org/10.16995/dm.30.
- Williams, George H. 2012. „Disability, Universal Design, and the Digital Humanities“. In Debates in the Digital Humanities, herausgegeben von Matthew K. Gold. Minneapolis. https://www.jstor.org/stable/10.5749/j.ctttv8hq.15.
- Wymer, Kathryn. 2005. „Why Universal Accessibility Should Matter to the Digital Medievalist“. Digital Medievalist, Nr. 1. https://doi.org/10.16995/dm.9.

One Document Does it all ... – Nutzerzentriertes Design inklusive

Quenouille, Nadine, Sächsische Akademie der Wissenschaften zu Leipzig

Lightning Talk

„Born-digital“-Editionen gelten heute als moderne Form der wissenschaftlichen Publikation. Dank des digitalen Formats können sie nahezu unbegrenzt Inhalte aufnehmen und jederzeit ergänzt oder erweitert werden. Gelungene Designs integrieren dabei eine Vielzahl an Informationen – vom Text über das Digitalisat bis hin zu Metadaten und ausführlichen Kommentaren – und machen sie so einer Vielzahl an interessierten Nutzern zugänglich.

Doch bisher wurden die technischen Möglichkeiten solcher Editionen oft noch nicht voll ausgeschöpft, so daß bestimmte Nutzergruppen – insbesondere sehbehinderte und sehbeeinträchtigte Personen – weiterhin Schwierigkeiten beim Zugang zu textlichen Dokumenten haben.

Im Rahmen des Langzeitvorhabens Forschungsportal BACH stehen wir kurz vor der Veröffentlichung unserer Webseite, die Ende des Jahres mit den ersten Editionen von Dokumenten rund um die Familie von Johann Sebastian Bach online gehen soll. Unsere Editionen enthalten eine Fülle an Informationen: von Personen- und Ortsdaten über Abkürzungen und deren Auflösungen bis hin zu zeilenspezifischen Kommentaren. Aus Platz- und Gestaltungsgründen kommen dabei auch Popovers zum Einsatz.

Doch eine solche Darstellung schließt die Gruppe der sehbeeinträchtigten und sehbehinderten Personen aus, da dünne, gepunktete Linien unter Wörtern mit Popovers für sie schwer erkennbar sind. Ästhetik spielt für diese Nutzergruppe keine Rolle, wohl aber Design: Design im Sinne von proaktiver Gestaltung, um auch ihnen Zugang zu historischen Texten mitsamt ihren reichhaltigen Informationen, die eine Edition bietet, zu geben.

Im Mittelpunkt dieses Lightning Talks steht nun die Verknüpfung einer klassischen Edition mit modernem Webdesign, um eine flexible und für viele verschiedene Bedürfnisse zugängliche Präsentation von Texteditionen zu ermöglichen. Dabei geht es nicht nur um Einhaltung von Standards wie WCAG-Richtlinien, sondern um ein Design, das Vielfalt als Ausgangspunkt nimmt. Features wie etwa Kontrastoptionen, anpassbare Schriftgrößen oder Dyslexie-Fonts werden nicht als nachträgliche Add-ons, sondern als integraler Bestandteil des Designs verstanden.

Die technische Grundlage bilden unsere Dokumente im TEI-Format, deren Darstellung im Web durch Javascript und ein ODD (One Document Does it all) unterstützt wird. Diese Technologien ermöglichen es, semantische Tiefe mit technischer Flexibilität zu verbinden. Nutzer können hierbei selbst entscheiden, wie die Darstellung der Edition ihren individuellen Bedürfnissen angepasst wird. Der Vorteil von TEI und ODD liegt hierbei nicht nur in der Interoperabilität, sondern vor allem in der

Flexibilität zur Gestaltung verschiedener Nutzungsszenarien. Barrierearmut wird hierbei nicht als nachträgliche Pflichtübung, sondern als kreativer Gestaltungsprozess begriffen.

Anhand eines Dokuments aus unserer Edition wird gezeigt, wie digitale Inklusion in der Praxis umgesetzt werden kann – nämlich durch die Kombination von Standards, Open-Source-Tools und nutzerzentriertem Design.

Neue Wege in die Reisetagebücher der *edition humboldt digital*

Stefan Dumont; Tobias Kraft – Berlin-Brandenburgische Akademie der Wissenschaften

Lightning Talk

Seit 2015 werden im Akademienvorhaben „Alexander von Humboldt auf Reisen – Wissenschaft aus der Bewegung“ die Reisetagebücher, Briefe und weitere Dokumente zu Humboldts Reisen nach Südamerika (1799–1804), Zentralasien (1829) und in Europa (z. B. nach England 1790) ediert und in der *edition humboldt digital* (*ehd*) (https://edition-humboldt.de) und *edition humboldt print* herausgegeben.

Für die aktuelle Version 12 der *edition humboldt digital*, die im Juni dieses Jahres erschien, wurden drei Funktionen entwickelt, die den Benutzer:innen neue Zugänge zu und Perspektiven auf die edierten Reisetagebücher bieten. Dreh- und Angelpunkt ist dabei die Chronologie zu A. v. Humboldts Leben, die bereits in den 1960er Jahren begonnen, in den 2000ern erstmals im Web publiziert und seit Projektbeginn im Rahmen der *ehd* veröffentlicht und kontinuierlich erweitert wird. Entwickelt wurden folgende drei neue Zugänge:

1. Auf einer Kartenansicht können die Stationen und ihre zeitliche Abfolge einer Forschungsreise von Humboldt visualisiert werden. Die dafür notwendigen Daten bezieht die Visualisierung nicht aus den edierten Texten selbst, sondern aus der AvH-Chronologie, da diese eine kohärente Datenbasis bietet. In den Chronologieeinträgen werden alle Aufenthaltsorte einer Reise Humboldts durch die Vorhabensmitarbeiter:innen entsprechend markiert, die Koordinaten werden über die im Ortsregister hinterlegte GeoNames-URI bezogen. Zu jeder Station wird in der Visualisierung der Chronologieeintrag des jeweiligen Tages angezeigt. Derzeit sind zwei Reisen visualisiert (England 1790 und Zentralasien 1829), weitere werden vom Vorhaben in Zukunft ergänzt.
2. Neben dem Tageseintrag zeigt die Visualisierung darüber hinaus auch die Stellen an, an denen das Datum des Tages in den edierten Reisetagebüchern (und anderen Texten) in der *ehd* erwähnt wird. Per Klick kann man so aus der Visualisierung zu den verschiedenen Textstellen springen, die auf diesen Tag Bezug nehmen. Das geht natürlich auch andersrum: von den in den Texten genannten Datumsangaben kann man über ein Popup in die Chronologie und die Reisevisualisierung gelangen. Dadurch werden erstmals die Reisetagebücher auch aus der Perspektive der Reise heraus zugänglich und nutzbar.
3. Die 1685 Einträge in der A. v. Humboldt-Chronologie wurden erstmals in Englisch übersetzt – und zwar automatisiert über die API des Translationstool DeepL. Die DeepL-API kann mit XML umgehen, so dass Verknüpfungen zu Registereinträgen auch in der übersetzten Version erhalten bleiben. Auf die automatisierte Übersetzung wird bei jedem Eintrag entsprechend hingewiesen. Durch diese Funktion soll ein zentraler Teil der *ehd* für die internationale Humboldt-Forschung noch stärker geöffnet werden.

Mit den drei oben vorgestellten neuen Funktionen der Version 12 der *edition humboldt digital* wurde die A. v. Humboldt-Chronologie so weiterentwickelt, dass sie nunmehr nicht nur einfach eine zusätzliche Informationsressource darstellt, sondern neue Wege in die edierten Reisetagebücher, Briefe und Dokumente ermöglicht und sie miteinander vernetzt.

Literatur

- Dumont, Stefan. 2026. „Digitale Methodik. Ein Überblick über Datenmodellierung, Technologien und Schnittstellen in der *edition humboldt digital*." In *edition humboldt digital*, herausgegeben von Ottmar Ette. Version 12, 9. Juni 2026. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. https://edition-humboldt.de/v12/H0016212.
- Dumont, Stefan, Sabine Seifert, Tobias Kraft, Christian Thomas und Jan Wierzoch. 2026. „Data on the Move." *Journal of the Text Encoding Initiative*, Nr. 18. http://journals.openedition.org/jtei/6215.
- Ette, Ottmar, Hrsg. 2026. *edition humboldt digital*. Version 12. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. https://edition-humboldt.de.
- Kraft, Tobias, und Stefan Dumont. 2020. „The Humboldt Code. On Creating a Hybrid Digital Scholarly Edition of a 19th Century Globetrotter." *Wiener Digitale Revue* 1. https://doi.org/10.25365/wdr-01-03-02.
- Schwarz, Ingo, Hrsg. 2026. Alexander von Humboldt-Chronologie. In *edition humboldt digital*, herausgegeben von Ottmar Ette. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. https://edition-humboldt.de/X0000001.

Session 2: Sprache, Wörterbücher & KI

Das Etymologika-Vorhaben: digitale Zugänglichmachung eines griechisch-byzantinischen Wörterbuchs

Daniel Deckers, Universität Hamburg / Akademie der Wissenschaften in Hamburg; Kajetan Dvoracek, Universität Hamburg / Akademie der Wissenschaften in Hamburg

Vortrag

Die griechisch-byzantinischen Etymologika sind Wörterbücher, deren Einträge die ‚wahre Bedeutung' (étymon) der Wörter erhellen sollten; sie zählen zu den bedeutendsten lexikographischen Zeugnissen antiker und mittelalterlicher Wissensgeschichte in Europa. Sie sind zugleich Zeugnisse byzantinischer Gelehrsamkeit und mit ihren Beispielen eine Quelle von Zitaten aus der antiken griechischen Literatur, die oft nicht anderweitig erhalten sind. Das „Etymologika"-Vorhaben der Akademie der Wissenschaften in Hamburg hat sich die erstmalige vollständige kritische Edition eines bedeutenden Etymologikon und dessen umfassende Erschließung (mit Übersetzung, Kommentar usw.) zum Ziel gesetzt. Das sogenannte Etymologicum Gudianum (aus der zweiten Hälfte des elften oder dem Anfang des zwölften nachchristlichen Jahrhunderts) weist die Besonderheit auf, dass die Handschrift, die das ursprüngliche Arbeitsexemplar seiner Erstellung darstellt, weitgehend erhalten ist. Zugleich kann – nicht nur mit Blick auf die im Arbeitsexemplar fehlenden oder beschädigten Textpartien – auf eine reiche auf diesem aufbauende Überlieferung, Weiternutzung und Ergänzung zurückgegriffen werden, die bis in die beginnende Renaissance reicht.

In diesem Vortrag wird insbesondere auf die Zugänglichmachung des Projektgegenstandes und der erstellten Materialien eingegangen; dabei stehen die verschiedenen Formen der Online-Präsentation, der zugrundeliegenden Daten und der ein- bzw. angebundenen Erschließungshilfen im Vordergrund. Daneben werden auch einige typische Fragestellungen, die sich im Umfeld digitalisierter Workflows in einem philologischen Projekt ergeben, angesprochen.

Die Zielgruppen unseres Vorhabens sind überwiegend ein wissenschaftliches Fachpublikum oder jedenfalls wissenschaftlich Interessierte; das Material und seine Überlieferung sind jedoch nicht ausschließlich für die Manuskriptforschung und die griechisch-byzantinische Lexikographie relevant (auch für diese Gruppen stellt die Form der handschriftlichen Überlieferung bereits eine Herausforderung dar), sondern zudem für zahlreiche Nachbardisziplinen, denen mit einer klassischen Edition ohne zusätzliches Erschließungsmaterial kaum geholfen wäre. In einem kurzen Überblick wird thematisiert, welche Erwartungen dieser weiter gefassten Nutzergruppen berücksichtigt wurden und wie diesen technisch und bei der Aufbereitung des Materials Rechnung getragen werden kann, während einige Kernbereiche Gegenstand der folgenden Einzeldarstellungen sind.

Im Projekt wird einerseits eine klassische gedruckte Edition erstellt, die neben dem kritischen Text mit Apparaten eine gegenübergestellte Übersetzung (in englischer Sprache) mit einem Kurzkommentar enthält. Andererseits bietet die parallel entstehende Online-Edition Raum für einen weit ausführlicheren Kommentar und ermöglicht die Einbindung zahlreicher zusätzlicher Ressourcen – sowohl solcher, die bereits von Dritten verfügbar sind, als auch selbst erstellter bzw. erfasster. Gegenstand dieser Präsentation wird vor allem die Online-Edition und deren Zugänglichkeit sein, weshalb entlang einer detaillierten Darstellung zur Methodik der Erschließung verschiedener Materialarten auch die Formen der Präsentation und die dahinterstehenden Überlegungen zur besseren Zugänglichmachung betrachtet werden sollen.

Zunächst wird hier auf die Einbindung bereits online verfügbarer, insbesondere offener bzw. im Open Access zugänglicher Daten eingegangen werden. Hierzu zählen zum einen Handschriftendigitalisate (insbesondere auch ein Volldigitalisat der Haupthandschrift), die über einen Viewer überwiegend direkt von den IIIF-Servern der Bibliotheken abgerufen werden, zum anderen sowohl bekannte griechische Texte (Quellen, Parallelen usw.), die zur Einbindung z.B. aus Perseus / der Scaife Digital Library abgerufen oder bei kostenpflichtigen Angeboten (TLG) nur verlinkt werden können, als auch lexikographische und grammatische Materialien, insbesondere die über das Alpheios Project verfügbar gemachten. Thematisiert werden neben den Datenformaten die Art der Anbindung bzw. Integration in die Oberfläche und der Verweise sowie die Einbettung in die eigene(n) Präsentationsumgebung(en) und die dabei teilweise erfolgende weitergehende Erschließung dieser Materialien.

Zentrales Anliegen der Online-Edition ist die Zugänglichmachung der eigenen Forschungsergebnisse und ihre Einbettung in die Begleitmaterialien. Die über die grundsätzlichen Prinzipien bei der Erstellung einer digitalen Edition hinausgehenden Anforderungen an kritische und Belegstellenapparate sowie an die Übersetzung und den Kommentar werden knapp skizziert. Als Besonderheiten kommen zunächst die weitergehende Erschließung der Handschriftendigitalisate hinzu, vor allem durch Annotation der einzelnen Lexikoneinträge in den Haupthandschriften, die Nutzern mit grundlegenden paläographischen Kenntnissen insbesondere den Zugang zu dem in seinem Layout und seiner Lesbarkeit herausfordernden Originalmanuskript erleichtert und bei der Heranziehung der übrigen Digitalisate eine Zeitersparnis bewirken kann. Daneben spielt die Ergänzung sprachlicher Ressourcen um Sondermaterial aus dem Umfeld der Etymologika und die Zugänglichmachung weiterer Quellen- und Parallelmaterialien, die bisher nicht online verfügbar sind, eine wichtige Rolle. Auch die Zusammenführung der jeweiligen bibliographischen Hinweise in verschiedenartigen Materialien, um nicht nur eine entsprechende Gesamtbibliographie, sondern die für den spezifischen Kontext relevanten Auswahlen bieten zu können, wird vorgestellt. Es werden sowohl die gewählten Präsentationsformen als auch die Einbettung und Verknüpfung der verschiedenen und verschiedenartigen Ressourcen thematisiert, weiterhin die jeweils dahinterstehenden Formate, die Erfassungs- bzw. Aufbereitungsmethoden und deren für die jeweilige Online-Nutzbarmachung gewählte Tiefe.

Die Online-Edition verfügt sowohl über eine Desktop-orientierte Web-Oberfläche, die die parallele Ansicht (fast) aller Ressourcen in relativ frei anpassbaren Layouts ermöglicht, als auch über eine auf Mobilgeräte ausgerichtete Fassung, die auf die Präsentation aller wesentlichen Elemente in einer konziseren Form abzielt, ohne dabei auf Anpassungsmöglichkeiten für verschiedene Nutzergruppen zu verzichten. Hinsichtlich der Zugänglichkeit werden insbesondere die Abwägungen zwischen möglichst ausführlichem Material, möglichst umfassenden Ansichten sowie weitgehender Freiheit der Nutzer in der Auswahl einerseits und einer auf die jeweils entscheidenden Aspekte und Details reduzierten Präsentation, die notwendigerweise weitergehende Annahmen des Forschungsteams über die Nutzungsszenarien festschreibt, andererseits diskutiert werden.

In Hinsicht auf technische und Datenhaltungs-Aspekte wird dargelegt, welche Überlegungen z.B. zur besseren Nachnutzbarkeit in Aspekte wie ein statisches Hosting der Webapp, exportierbare und neue Formate, die Übernahme von Details der Print-Edition zur einheitlichen Zitierbarkeit auch aus der Online-Edition usw. eingeflossen sind. In Verbindung mit der Nachnutzbarkeit der verwendeten Formate wird zudem ein für ein allgemeineres Publikum erprobter Wiki-Export vorgestellt.

Abschließend wird kurz auf den experimentellen Einsatz von KI eingegangen, im Gegensatz zu den Kernpunkten des Vortrags jedoch eher in der Form eines perspektivischen ‚Laborberichts'. Auf der Ebene der Web-Oberfläche haben wir als Ergänzung zur klassischen Suche eine KI-gestützte Suchfunktion (auf Basis eines offenen Modells) getestet sowie die Bereitstellung der Suchdaten als MCP-Service. Außerdem berichten wir kurz über den Einsatz aktueller Modelle sowohl für das Korrekturlesen englischsprachiger Fachtexte als auch zur Identifizierung von Fehler- bzw. möglichen Problemstellen in der Übersetzung unseres Quellentexts. Dabei ist auch die Verwendung von TEI-Formaten für den Datenaustausch mit dem LLM-Thema, insbesondere die Vorteile des XML gegenüber reinem Text oder Office-Dokumenten zur Beibehaltung von Strukturen möglichst ohne negative Effekte auf die Verarbeitung durch die KI-Systeme. Zugleich beschreiben wir unsere Workflows mit der bidirektionalen Umwandlung in Textdokumente mit Anmerkungen bzw. Änderungen, um die Review-Prozesse insbesondere auf fachwissenschaftlicher Seite zu vereinfachen.

Literatur

- Alpers, Klaus. 1990. „Griechische Lexikographie in Antike und Mittelalter. Dargestellt an ausgewählten Beispielen." In Welt der Information. Wissen und Wissensvermittlung in Geschichte und Gegenwart, edited by H. A. Koch and A. Krup-Ebert. J. B. Metzler.
- Brockmann, Christian, José Maksimczuk, and Eva Wöckener-Gade, eds. Forthcoming 2026. The manuscript transmission of the Etymologicum Gudianum. Etymologika-Studien, 4. Vandenhoeck & Ruprecht.
- Deckers, Daniel, Kajetan Dvoracek, and Stefano Valente. Forthcoming. „The Etymologika Project: Editing a Medieval Greek Lexicon in Digital and Print Format." In Proceedings of the Conference *Fontes sine quibus non*: Interdisciplinary Approaches on Critical Editing, Text Encoding, and Text Mining (Athens, June 2023).
- Reitzenstein, Richard. 1897. Geschichte der griechischen Etymologika. Ein Beitrag zur Geschichte der Philologie in Alexandria und Byzanz. Teubner.
- The Linux Foundation. 2026. „What is the Model Context Protocol (MCP)?" Last modified July 28. https://modelcontextprotocol.io/.

Eine neue Mittelniederdeutsche Grammatik

Amy Isard, Universität Hamburg; Lena Schnee, Universität Hamburg; Chris Biemann, Universität Hamburg

Lightning Talk

Das neue Langzeitvorhaben Mittelniederdeutsche Grammatik (https://www.mingra.uni-hamburg.de/) (MiNGra) der Akademie der Wissenschaften in Hamburg hat als Ziel, eine neue korpusbasierte wissenschaftliche Grammatik des Mittelniederdeutschen zu produzieren. Die Grammatik hat als Basis das Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650) (https://www.slm.uni-hamburg.de/ren.html) (ReN) (Barteld et al., 2017; Schröder et al., 2017) das nach Wortart und Flexionsmorphologie annotiert ist. Das an der Universität Hamburg angesiedelte Projekt hat im Januar 2026 begonnen und läuft für 18 Jahre bis Dezember 2043.

Die Grammatik wird nicht nur als traditionelles Buch in mehreren Bänden erscheinen, sondern auch auf einer interaktiven Webseite präsentiert. Wir werden eine Online-Version der Grammatik bereitstellen, die auf die Bedarfe mehrerer Nutzergruppen (Experten, Studierende, Laien) zugeschnitten sein wird. Zudem wird den Nutzenden ermöglichen werden, über ein Suchportal mit der Grammatik, den zugehörigen Textquellen und weiteren Forschungsdaten zu interagieren, sowie eigene quantitative Visualisierungen der Ergebnisse zu erstellen. Erreicht werden diese Ziele durch die Verbindung linguistischer und informatischer Expertisen im Projektrahmen.

Die Basis für das Projekt ist das ReN-Korpus. Im Mittelpunkt steht die Untersuchung grammatischer Variation. Unterschiede können durch Raum, Zeit oder Textsorte bedingt sein und werden systematisch erfasst. In unserem Workflow werden die Korpustexte für die Grammatik analysiert, wobei grammatikrelevante Annotationen zum Korpus hinzugefügt werden mitsamt gelegentlicher Korrekturen der bestehenden Annotationen. Eine Instanz des Such- und Visualisierungs-Tools ANNIS (https://corpus-tools.org/annis) (Krause und Zeldes, 2016) für das ReN-Korpus ist bereits vorhanden. Basierend auf der Arbeit des Projektes werden in Zukunft neue Versionen des ReN-Korpus veröffentlicht werden, die neue Annotationen sowie Korrekturen bereitstellen.

Die geplante Grammatik wird zahlreiche Belege aus dem ReN-Korpus enthalten, die in der Online-Fassung Hyperlinks zu den jeweiligen Textstellen der ReN-ANNIS-Instanz erhalten. Zusätzlich wird auf Wortebene zu Einträgen des Mittelniederdeutschen Wörterbuch (https://woerterbuchnetz.de/?sigle=MNWB&lemid=A00001) verlinkt, das im Rahmen des Wörterbuchnetzes online verfügbar ist.

Um den Schreib- und Produktionsprozess zu optimieren, planen wir, die Texte für die Grammatik mit Pandoc (https://pandoc.org) zu verfassen. Auf diese Weise kann die Grammatik als Buch in Druck- und PDF-Fassung, sowie als digitale Edition in Form einer Webseite im HTML-Format veröffentlicht werden. Da das Langzeitprojekt eine Laufzeit von 18 Jahren hat, können wir so auch bei zukünftige Entwicklungen im Bereich der Webtechnologien flexibel reagieren und Inhalte in verschiedenen Formaten bereitstellen. Diese Vorgehenweise ermöglicht es den Linguist:innen, sich auf die wissenschaftliche Basis der Grammatik zu konzentrieren, während beide Gruppen kollaborativ das Buchlayout und interaktive Webseite konzipieren, die von den Informatiker:innen implementiert wird.

Wir entwickeln ein Konzept in dem Pandoc und LaTeX kombiniert werden, um den Schreibprozess so unkompliziert wie möglich zu machen. Hierzu haben wir beispielsweise ein Format für die Referenzierung der Beispielbelege aus dem Korpus gewählt, welches in der Eingabe menschen- und maschinenlesbar ist und in der Ausgabe in mediengerechter Form realisiert wird: In der digitalen Edition als reiner Textverweis (Textname, Token-ID) und in der Online-Grammatik als Hyperlink zur

Korpusquelle. Die Bibliographie wird auf Grundlage einer Zotero (https://www.zotero.org)-Bibliothek erstellt und kann so ebenfalls in verschiedenen Formaten ausgegeben werden. Kapitel, Abschnitte, Beispiele, Tabellen, Abbildungen usw. werden in Markdown-Syntax notiert, und IDs werden automatisch generiert, um Referenzen innerhalb des Textes zu ermöglichen. Wichtig bei diesem Vorgehen ist, Markup-Notationen zu wählen, die eindeutig und ausdrucksstark genug für die automatische Verarbeitung sind, aber gleichzeitig einfach und komfortabel genug in der Nutzung von weniger technikorientierten Projektmitgliedern.

Wir werden eine Overleaf Instanz der Gesellschaft für wissenschaftliche Datenverarbeitung (https://gwdg.de) (GWDG) benutzen, die eine kooperative Arbeitsumgebung zur Vefügung stellt, bei der mehrere Forschende gleichzeitig auf demselben Text arbeiten können. Da wir unsere Markdown-Dokumente in einem LaTex-Dokument integrieren, profitieren wir von dort verfügbaren Features wie der Autovervollständigung für Bibliographieeinträge. Jedes Overleaf-Dokument ist gleichzeitig ein Git-Repositorium. Neben der daraus resultierenden automatischen Versionshistorie erlaubt dies auch die Integration in integrierte Workflows, da der aktuelle Stand eines Dokuments jederzeit exportiert werden und nach Bearbeitung (z.B. Generierung von IDs) erneut importiert werden kann.

Im Sinne der Nachhaltigkeit der Projektergebnisse soll die Entwicklung und Veröffentlichung der Grammatik des Mittelniederdeutschen auch möglichst weitreichend der Praxis der offenen Forschung folgen. Die Nutzung offener etablierter Standards wie Markdown und die Nutzung offener Software wie Pandoc, LaTeX und Zotero ist hierzu der erste Schritt. Auch Daten und Code des Projekts sollen möglichst offen zur Verfügung gestellt und dokumentiert werden. Das Forschungsdatenmanagement der Universität Hamburg erlaubt es dem Projekt hierbei, seine Forschungsergebnisse den FAIR-Prinzipien (Wilkinson et al., 2016) entsprechend zu publizieren und archivieren, um sie langfristig zu erhalten.

Literatur

- Barteld, Fabian, Katharina Dreessen, Sarah Ihden, und Ingrid Schröder. 2017. Das Referenzkorpus Mittelniederdeutsch / Niederrheinisch (1200-1650) - Korpusdesign, Korpuserstellung und Korpusnutzung (https://doi.org/10.14220/mdge.2017.64.3.226). Anja Becker und Albrecht Hausmann (Hrsg.): Mittelniederdeutsche Literatur. Mitteilungen des deutschen Germanistenverbandes 64. Jg./H. 3, Göttingen: Vandenhoeck & Ruprecht, S. 226–241.
- Krause, Thomas, und Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization (https://doi.org/10.1093/llc/fqu057). Digital Scholarship in the Humanities, 31(1), S. 118–139.
- Schröder, Ingrid, Fabian Barteld, Katharina Dreessen, und Sarah Ihden. 2017. Historische Sprachdaten als Herausforderung für die manuelle und automatische Annotation: Das Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200-1650) (https://www.slm.uni-hamburg.de/ren/media/pdf/sd-schroeder.pdf). Jahrbuch des Vereins für niederdeutsche Sprachforschung 140, S. 43-57.
- Wilkinson, Mark, Michel Dumontier, I Jsbrand Jan Aalbersberg et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship (https://doi.org/10.1038/sdata.2016.18). Scientific Data, 3:160018.

Wie lassen sich KI-Tools für kleine nordeurasische Sprachen einsetzen?

Elena Lazarenko, Universität Hamburg; Aleksandr Riaposov, Universität Hamburg; Mariya Brykina, Universität Hamburg; Beáta Wagner-Nagy, Universität Hamburg

Lightning Talk

In dieser Präsentation werden die Einsatzbereiche verschiedener KI-Tools im Korpuserstellungs-Workflow des INEL-Projekts dargestellt. Das Projekt INEL ist ein Langzeitvorhaben, das darauf abzielt umfassend annotierte digitale multimodale Korpora indigener, bedrohten Sprachen Nordeurasiens zusammenzustellen und zu veröffentlichen. Die Quelldaten für die Korpora sind unterschiedlich: Feldforschungsmaterialien (Aufnahmen sowie ihre Transkriptionen), handschriftliche Manuskripte oder bereits veröffentlichte Texte. Im Rahmen des Projektes wurden bislang neun Korpora veröffentlicht.

Während viele sprachwissenschaftliche Projekte KI-gestützte Lösungen für die automatische Annotation und Übersetzung von Daten nutzen, stellen die Low-Ressource-Sprachen nach wie vor eine große Herausforderung dar. Trotz der rasanten Fortschritte der letzten Jahre, haben die LLMs nach wie vor gewisse Schwierigkeiten bei der Analyse ressourcenarmen Sprachen. Oft ist die Erhebung von zusätzlichen Daten erforderlich (vgl. Ginn et al. 2024a; Almendra et al. 2026; Kipyatkova et al. 2025; Pava et al. 2025; Zhong et al. 2025).

Dies gilt insbesondere für die Erstellung von Transkripten mit interlineare Morphemglossierung (auch IGTs, engl. „Interlinear Glossed Text“). Ein Beispiel für eine solche Glossierung ist in (1) dargestellt.

(1) Nordmansisch

*am	ti	mars’əm-awä-s-ə-m*
PRO1SG	DISC	be.bored-PASS-PST-EP-1SG.S

„Mir wurde es schon langweilig.“

VIYa_18890125_CreationOfEarth_flk.248 (INEL / Munkácsi 1902: 1-31)

Bei der Glossierung handelt es sich um zwei Aufgaben: erst wird jedes Wort in Teile (oder Morpheme) segmentiert, dann bekommt jedes Teil ein passendes Kennzeichen wie PASS (Passivmarker) und PST (Past). Mithilfe von Programmen wie SIL FLEx (https://software.sil.org/fieldworks/), die über einen eingebauten morphologischen Parser verfügen, verläuft der Prozess zwar schneller, halbautomatisiert, bleibt aber zeitaufwändig.

Obwohl es State-of-the-Art-Modelle zur IGT-Generierung und Verfahren, mit denen sich die Low-Ressource-Sprachen wirksam mithilfe von LLMs übersetzen lassen, gibt, reichen sie noch nicht aus, um diese Aufgabe zuverlässig und vollständig zu automatisieren (vgl. Ginn et al. 2024b; Shandilya et al. 2026; Zhang et al. 2024; Rice et al. 2025).

Im Fall von INEL werden die beiden Aufgaben (automatische Glossierung und Übersetzung) noch komplizierter: die Sprachen sind nicht standardisiert und die Schreibweise und die Transkriptionskonventionen unterscheiden sich stark von Quelle zu Quelle.

Das automatische Transkribieren von Low-Ressource-Sprachen ist ebenfalls möglich, jedoch häufig fehlerhaft. Dafür wird ein gut vorbereitetes Dataset mit ausreichend vielen und qualitativ geeigneten Daten benötigt, um ein neues Modell zu trainieren, oder das vorhandene Modell an eine bestimmte

Sprache anzupassen (vgl. Partanen et al. 2020; Taguchi et al. 2026). Und so schließt sich der Kreis: Um KI bei der Korpuserstellung einzusetzen, braucht man oft zuerst ein Korpus. Oder sogar: egal wie fortgeschritten ein Modell ist, wenn es nicht mit genügend geeigneten Sprachdaten versorgt wird, wird es wahrscheinlich underperformen.

Da die Implementierung von KI-basierten selbstentwickelten Schnittstellen im Rahmen des INEL-Projekts derzeit weder zeitlich noch methodisch sinnvoll ist, erfolgen die grammatikalischen Analysen weiterhin überwiegend manuell. KI-basierte Tools werden dabei nicht vollständig aus unserem Workflow ausgeschlossen, sondern so eingesetzt, dass der Zeitaufwand bei der Korpuserstellung verringert wird, ohne die Korpusqualität zu beeinträchtigen:

1. Optical Character Recognition von Manuskripten und Büchern. Transkribus hat sich als nützliches Werkzeug für OCR- (Optical Character Recognition) und HTR- (Handwritten Text Recognition) Aufgaben erwiesen. INEL nutzt Transkribus, um die Vorbereitung von Textdaten zu beschleunigen. Unter anderem wurden zwei Modelle dabei veröffentlicht. Das erste wurde anhand handgeschriebenen Ewenkisch-Russischen Manuskripten von Konstantin Rychkov aus den 1910s trainiert. Das zweite Modell wurde für die gedruckte Finno-Ugrische Transkription von Tundra- und Waldnenzischen Texten mit über 100 einzigartigen Zeichen und Diakritika entwickelt.

2. Für die meisten Texte, die Teil eines INEL-Korpus werden, ist bereits eine ursprüngliche (meistens russische) Übersetzung in der Quelle vorhanden. Die Sprache dieser Übersetzung sowie deren Wohlgeformtheit variieren je nach Quelle – z. B., bedeutet das häufig für ältere Quellen, dass die Übersetzung vor der Veröffentlichung an die heute gültige Schreibweise angepasst werden soll. Außerdem werden die veröffentlichten Texte in der Regel ins Englische, Deutsche und Russische, sofern entsprechende Übersetzungen nicht bereits vorhanden sind, übersetzt. Obwohl eine zuverlässige KI-gestützte Übersetzung direkt aus den Quellsprachen in unserem Fall kaum möglich ist, ist die Übertragung zwischen den „großen“ Sprachen dank KI erheblich schneller geworden. Wir nutzen das DeepL API, um die vorläufige Übersetzung bereitzustellen. Diese Übersetzungen erfordern zwar häufig eine manuelle Nachbearbeitung, verkürzen jedoch die Bearbeitungszeit bei der Erstellung der Korpora deutlich.

Literatur

- Almendra, Anaís, Arianna Bisazza, Claudio Gutierrez, und Felipe Hasler. 2026. „What Resources Matter for Interlinear Glossing? Using LLMs and RAG for the Low-Resource Mapudungun Language“. Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), 64–73.
- Ginn, Michael, Mans Hulden, und Alexis Palmer. 2024. Can we teach language models to gloss endangered languages? Übersetzt von Yaser Al-Onaizan, Mohit Bansal, und Yun-Nung Chen. Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics. https://aclanthology.org/2024.findings-emnlp.337/, https://doi.org/10.18653/v1/2024.findings-emnlp.337.
- Ginn, Michael, Lindia Tjuatja, Taiqi He, u. a. 2024. GlossLM: A Massively Multilingual Corpus and Pretrained Model for Interlinear Glossed Text. Übersetzt von Yaser Al-Onaizan, Mohit Bansal, und Yun-Nung Chen. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.683/, https://doi.org/10.18653/v1/2024.emnlp-main.683.
- Kipyatkova, Irina, Ildar Kagirov, und Mikhail Dolgushin. 2025. „Use of Pre-Trained Multilingual Models for Karelian Speech Recognition“. Informatics and Automation 24 (2): 604–30. https://doi.org/10.15622/ia.24.2.9.
- Partanen, Niko, Mika Hämäläinen, und Tiina Klooster. 2020. „Speech recognition for endangered and extinct Samoyedic languages“. Proceedings of the 34th Pacific Asia conference on language, information and computation, 523–33.

- Pava, Juan N., Caroline Meinhardt, Haifa Badi Uz Zaman, u. a. 2025. Mind the (Language) Gap.
- Rice, Enora, Katharina von der Wense, und Alexis Palmer. 2025. „Interdisciplinary research in conversation: A case study in computational morphology for language documentation“. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 11284–96.
- Shandilya, Bhargav, Matt Buchholz, und Alexis Palmer. 2026. „GlossAssist -- A Tool to Simplify Corpus Creation and Study the Effect of NLP Models in Low-Resource Documentation Settings“. arXiv:2606.04367. Preprint, arXiv, Juni 3. https://doi.org/10.48550/arXiv.2606.04367.
- Taguchi, Chihiro, Yukinori Takubo, und David Chiang. 2026. „Automatic Speech Recognition for Documenting Endangered Languages: Case Study of Ikema Miyakoan“. arXiv:2603.26248. Preprint, arXiv, April 20. https://doi.org/10.48550/arXiv.2603.26248.
- Wagner-Nagy, Beáta, und Katalin Sipőcz. 2026. „INEL Mansi Corpus“. Version 1.0. Universität Hamburg, Juni 5. https://doi.org/10.25592/UHHFDM.18715.
- Zhang, Kexun, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, und Lei Li. 2024. „Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions“. arXiv:2402.18025. Preprint, arXiv, November 11. https://doi.org/10.48550/arXiv.2402.18025.
- Zhong, Tianyang, Zhenyuan Yang, Zhengliang Liu, u. a. 2026. „Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research“. arXiv:2412.04497. Preprint, arXiv, April 17. https://doi.org/10.48550/arXiv.2412.04497.

EdiKILex. Edition, Künstliche Intelligenz und Lexikographie als Schlüssel zu vormodernen Textwelten

Anja Lobenstein-Reichmann, NADW Göttingen; Lena Vosding, NADW Göttingen; Jan Christian Schaffert, NADW Göttingen

Lightning Talk

Der Zugang zu historischen Texten wird durch die Bereitstellung wissenschaftlicher Editionen und entsprechender Wörterbücher maßgeblich erleichtert. Sowohl Editionen als auch Wörterbücher sind zunehmend als digitale Daten verfügbar und werden digital erstellt. Diese Datensätze sind aber noch nicht vernetzt. Das Projekt EdiKILex begegnet diesem Desiderat am Beispiel des Frühneuhochdeutschen, für das neben dem FWB auch Editionen digital vorliegen, und entwickelt eine künstliche Intelligenz, die auf Wortebene interpretierende Hilfestellung gibt. Zugleich untersucht EdiKILex, welche Voraussetzungen für eine intuitive Benutzung gegeben sein müssen und welche Folgen KI-generierte Informationen für menschliche und maschinelle Rezeption haben.

Literatur

- Lobenstein-Reichmann, Anja, Oskar Reichmann / 2020. „Lexikographische Konzeption und kulturgeschichtliche Erkenntnis im digitalen Zeitalter.„ In: Metalexicography, Dictionaries and Culture, edited by Rufus H. Gouws, 287–309. Lexicographica. Internationales Jahrbuch für Lexikographie 36. Berlin/Boston: De Gruyter.
- https://adw-goe.de/forschung/weitere-forschungsprojekte/edikilex/.

Offline für alle: Die DWDS-App

Frank Wiegand, Digitales Wörterbuch der deutschen Sprache (DWDS)

Lightning Talk

Das Projekt „Digitales Wörterbuch der deutschen Sprache“ (https://www.dwds.de/) (DWDS) ist seit über zwei Jahrzehnten eines der zentralen lexikalischen Informationssysteme zur deutschen Gegenwartssprache und wird an der Berlin-Brandenburgischen Akademie der Wissenschaften

(https://www.bbaw.de/) (BBAW) betrieben (Klein & Geyken 2010). Seit November 2024 (Android[1]) beziehungsweise Februar 2025 (iOS[2]) steht das DWDS-Wörterbuch auch als App (https://www.dwds.de/a), die vollständig offline nutzbar ist, für mobile Endgeräte zur Verfügung. Im Vortrag soll über die Genese dieser Anwendung gesprochen werden. Dazu werden Anregungen gegeben, was Zugänglichkeit in den Digital Humanities jenseits reiner Onlineverfügbarkeit bedeuten kann. Technische Nutzbarkeit und Anschlussfähigkeit sind dabei ebenso gemeint wie freie Zugänglichkeit im engeren Sinne.

Die Vorarbeiten begannen im Februar 2023 mit der Frage, wie das DWDS auch Menschen ohne verlässliche bzw. erschwingliche Internetverbindung erreichen kann. Wie der D21-Digital-Index 2024/2025 zeigt, ist reine Konnektivität dabei nur ein Teil des Problems: In Deutschland zählen 4,2 Millionen Menschen zu den sogenannten Offliner:innen. Dabei ist der mit Abstand häufigste Grund für Nichtnutzung des Internets inzwischen nicht mehr fehlendes Interesse, sondern empfundene Komplexität digitaler Anwendungen (Initiative D21 2025). Eine bewusst reduzierte, offline funktionierende Wörterbuch-App adressiert damit eine Zugangsbarriere, die eher mit Einfachheit als mit Bandbreite zu tun hat. Als weitere Zielgruppe sollten Personengruppen adressiert werden, die aufgrund organisatorischer Einschränkungen ohne Netzzugang ein digitales Wörterbuch benötigen, also beispielsweise Schüler:innen oder Angestellte in abgesicherten IT-Umgebungen. Anstatt eine eigene App-Infrastruktur von Grund auf aufzubauen, entschied sich die DWDS-Projektleitung für eine Kooperation mit Kiwix, einer gemeinnützigen Organisation (Schweiz), die seit 2007 Open-Source-Software zur Offlinebereitstellung von Webinhalten entwickelt und Projekte wie Wikipedia oder Project Gutenberg verfügbar macht (Lin 2015; Coillet-Matillon 2020). Es wurde also eine bestehende, offene Infrastruktur nachgenutzt.

Für die Distribution der DWDS-Inhalte aus technischer Sicht werden dafür das DWDS-Wörterbuch sowie ein reduziertes Gerüst der Webseite www.dwds.de vollständig statisch als HTML generiert und in einen ZIM-Container verpackt. ZIM (ZIM file format 2026) ist ein offenes Kompressionsdatenformat, das Kiwix zur Offline-Distribution nutzt. Die DWDS-Inhalte werden über einen angepassten Viewer, der den Datencontainer ausliest, bereitgestellt. Die App selbst wird über Kiwix' Open-Source-Infrastruktur auf GitHub für „gebrandete" Apps gebaut und gepflegt, wodurch das DWDS von Kiwix' Erfahrungen mit Appstore-Anforderungen und plattformübergreifender Auslieferung profitiert, ohne selbst eine Codebasis unterhalten zu müssen.

Mit der Idee für die App offenbarten sich Herausforderungen für eine wissenschaftliche Organisationsstruktur wie die BBAW und das DWDS als befristetes Vorhaben: Die Registrierung als App-Anbieter, die Hinterlegung von Kreditkartendaten und Telefonnummern für Verifizierungsprozesse sowie der intransparente Review-Prozess bei Apple und Google waren hier ungewohnte bürokratische Hürden. So zeigte sich eine weitere Dimension von Zugänglichkeit, die im Diskurs oft weniger beleuchtet wird: Nicht nur Nutzer:innen stoßen auf technische und institutionelle Hürden, auch Anbieter offener Wissensressourcen müssen sich durch die Infrastrukturlogik kommerzieller Plattformen hindurcharbeiten, um überhaupt sichtbar zu werden.

Die Annahme der App bestätigt, dass sich dieser Aufwand lohnt: Die Stores von Apple und Google verzeichnen Bewertungen von rund 4,9 von 5 Sternen (bei ca. 700 Reviews), und Nutzer:innen heben in Rezensionen insbesondere die wissenschaftliche Verlässlichkeit und die Offline-Verfügbarkeit hervor, mehrfach auch als Ersatz für die (inzwischen eingestellte) Duden-App. Mittlerweile wurde die DWDS-App ca. 190.000 mal heruntergeladen. Neben dem Google Play Store und Apple App Store distribuiert das DWDS die ZIM-Daten auch frei, sie stehen damit auch außerhalb der Ökosysteme der großen Techfirmen zur Verfügung.

Die DWDS-App operationalisiert Zugänglichkeit gleich in mehreren Facetten: technisch, durch Offline-Nutzung unabhängig von Konnektivität; institutionell, durch Nachnutzung offener Infrastrukturen statt Ressourcenkonzentration; und ökonomisch, pragmatisch, dem Leitbild freier Forschung verschrieben: durch kostenlose, werbefreie Bereitstellung. Das Projekt macht sichtbar, dass Offenheit nach außen oft institutionelle Übersetzungsarbeit nach innen erfordert, zwischen der Logik von Forschungssoftware und den Prozessen kommerzieller App-Stores. Nicht zuletzt eröffnen sich damit auch dauerhaft Ressourcen bindende Aufgaben, was die Pflege, Aktualisierung und Rückbindung durch die Nutzenden anbelangt.

Endnoten

[1] https://play.google.com/store/apps/details?id=org.kiwix.kiwixcustomdwds&hl=en-US&ah=-elbd_d25uJY9yw7zB9ALzJO_iw
[2] https://apps.apple.com/de/app/dwds/id6741328425

Literatur

- Coillet-Matillon, Stéphane. 2020. „No Internet, No Problem.“ in Wikipedia @ 20: Stories of an Incomplete Revolution, ed. Joseph Reagle and Jackie Koerner. The MIT Press. https://doi.org/10.7551/mitpress/12366.003.0023
- DWDS. „Das DWDS als App“. Accessed on September 4, 2026. https://www.dwds.de/a
- Initiative D21 e. V. 2025. „D21-Digital-Index 2024/2025 – Jährliches Lagebild zur Digitalen Gesellschaft“. March. Accessed on September 4, 2026. https://initiatived21.de/publikationen/d21-digital-index
- Klein, Wolfgang, and Alexander Geyken. 2010. “Das digitale Wörterbuch der deutschen Sprache (DWDS).“ in Lexicographica 26 (2010): 79–96. https://doi.org/10.1515/9783110223231.1.79
- Lin, Jimmy. 2015. „The Sum of All Human Knowledge in Your Pocket: Full-Text Searchable Wikipedia on a Raspberry Pi.“ in Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (New York, NY, USA), JCDL ’15, June 21, 85–86. https://doi.org/10.1145/2756406.2756938
- ZIM file format. 2026. „ZIM file format – openZIM“. July 10. Accessed on September 4, 2026. https://wiki.openzim.org/w/index.php?title=ZIM_file_format&oldid=31493

Session 3: Digitalisierung & Quellenerschließung

Wenn Quellen flüstern. Zugänge zu beschädigten Quellen mittels Automated Text Recognition (ATR)

Petra Heinicker, Berlin-Brandenburgische Akademie der Wissenschaften; Frederik Skidzun, Akademie der Wissenschaften und der Literatur Mainz

Vortrag

Ein zentrales Problem bei der Erschließung historischer Quellen sind materielle Schäden, da sie den Zugang zum Forschungsgegenstand selbst erschweren. In unserem Beitrag wollen wir am Beispiel eines so genannten „Formelbuches“ aus dem 15. Jahrhundert Wege und Möglichkeiten aufzeigen, wie Anwendungen aus dem Bereich des maschinellen Lernens dabei helfen können, die Lesbarkeit einer durch Wasserschaden beschädigten Textquelle zu verbessern, und gleichzeitig die Grenzen dieser Methoden diskutieren. Wir erproben ein Verfahren aus dem Bereich der automatisierten Texterkennung (Automated Text Recognition, ATR), um vom Wasserschaden betroffene

Textabschnitte lesbar zu machen. Dieser Ansatz ist experimentell und reiht sich zwischen fotografischen Methoden zur Wiederherstellung unlesbar gewordener Textstellen und diffusionsbasierten „Reparaturmethoden“ ein. Wir stellen ein speziell zu diesem Zweck trainiertes Texterkennungsmodell vor und erörtern, welchen Beitrag es dabei leisten kann, eine durch Wasser beschädigte Textquelle wiederherzustellen. Zudem diskutieren wir Möglichkeiten, dieses Modell sinnvoll zu evaluieren.

Digitale und analoge Methoden verbinden Im Rahmen des Akademievorhabens Regesta Imperii - Regesten Kaiser Friedrichs III. (BBAW) untersuchen wir zurzeit ein so genanntes „Formelbuch“ aus dem 15. Jahrhundert. Diese im Herzogtum Österreich entstandene Handschrift versammelt auf über 800 Seiten mehrere hundert Formulare, die als Vorlagen zum Ausstellen von Urkunden dienten. Die Handschrift ist systematisch angelegt, die Urkunden betreffen verschiedenste Bereiche des weltlichen Rechts. Selbstzeugnisse, die Hinweise auf die Autorschaft bzw. auf mögliche Auftraggeber geben könnten, fehlen. Die Erschließung der Handschrift wird darüber hinaus aus mehreren Gründen erschwert. Zum einen sind die Urkundenformulare typischerweise undatiert und die Personen- und Ortsbezeichnungen vielfach anonymisiert. In vielen Fällen sind Eigennamen und Datumsangaben sogar bewusst verfremdet, was eine zeitliche und inhaltliche Einordnung zusätzlich erschwert. Schließlich ist die Handschrift von einem großflächigen Wasserschaden betroffen, durch welchen die Schrift auf mehr als einem Drittel der Seiten stark und auf zahlreichen weiteren Seiten in geringerem Umfang beschädigt wurde. Für den Umgang mit dem Wasserschaden bieten sich digitale Verfahren an, um einen ersten Zugang zu ermöglichen. Ausgangspunkt für alle digitalen Erschließungsansätze im Formelbuchprojekt ist eine eigens konzipierte Pipeline zur automatisierten Texterkennung (Automated Text Recognition, ATR). Diese basiert auf der integrierten Transkriptionsumgebung eScriptorium und der ATR-Engine Kraken (vgl. Kiessling 2019 sowie Stokes et al. 2021) und ermöglicht es, die mithilfe dieser Tools erstellten Transkriptionen anschließend noch weiter zu bearbeiten, um sie für unterschiedliche nachfolgende Analyseschritte vorzubereiten.

Lässt sich ein Wasserschaden lesen? Beschädigte Textträger stellen ein ‚Alltagsproblem‘ historischer Forschung dar, wobei sich naturgemäß zwei Problemfelder unterscheiden lassen: Ein Verlust von Beschreibmaterial, etwa durch Tierfraß, Fäulnis und ähnliche Fälle, oder ein Textverlust bei noch weitgehend intaktem Beschreibstoff, wie er etwa dann auftritt, wenn der Textträger Feuchtigkeit ausgesetzt war, und die Tinte in der Folge verschwommen oder ausgewaschen ist. Auf unseren Fall trifft vor allem letzteres zu. Der Wasserschaden hat auf allen der über 800 Seiten der Handschrift Spuren hinterlassen. Im ersten Drittel der Handschrift ist mitunter bis zur Hälfte der Textzeilen einer Seite von den Auswaschungen betroffen, wobei sich unterschiedliche Beschädigungsgrade unterscheiden lassen. Teils ist die Tinte lediglich leicht verlaufen, wobei sich die einzelnen Buchstaben noch gut lesen lassen. An anderen Stellen sind Zeichen stärker erodiert und lassen sich nur noch teilweise erkennen. Schließlich gibt es Bereiche, die ganz ausgewaschen sind, sodass sich selbst bei genauem Hinsehen zumeist keine Buchstaben mehr erkennen lassen. Diese unterschiedlichen Beschädigungsgrade treten meist gemeinsam innerhalb einzelner Zeilen auf. Es existieren bereits Verfahren, um Textverlust in historischen Texten zu begegnen. Eine Möglichkeit stellen spezialisierte Fotografietechniken dar, insbesondere die sogenannte Multispektralfotografie (Multispectral imaging, MSI). Mithilfe einer Spezialkamera wird dabei Licht im Infrarot- sowie Ultaviolettbereich aufgenommen, wodurch für das menschliche Auge nicht mehr sichtbare Schriftreste dargestellt werden können. Auch Computer Vision-Verfahren lassen sich einsetzen, um verlorene Textsequenzen digital zu ‚reparieren‘. Für unser Vorhaben stellen allerdings beide Vorgehensweisen keine idealen Lösungsansätze dar: Die MSI wurde zwar bereits bei ähnlichen Problemfällen erfolgreich genutzt (vgl. DHInfra.at 2026), allerdings sind die Anschaffungskosten der dafür notwendigen Ausrüstung

sehr hoch. Anwendungen, welche die beschädigten Textbereiche digital wieder herstellen sollen, scheiden in unseren Überlegungen aus einem anderen Grund aus. Sie stellen die betroffenen Abschnitte nicht wieder her, sondern erstellen lediglich eine probabilistische Vorhersage, wie diese ursprünglich ausgesehen haben könnten (vgl. Yang et al. 2025, S. 7). Für unseren Anwendungsfall erproben wir daher einen anderen Ansatz. Dafür nutzen wir die ATR-Pipeline, welche den Ausgangspunkt aller digitalen Ansätze im Formelbuchprojekt bildet. Wir trainieren ein Kraken-Texterkennungsmodell, welches es ermöglichen soll, auch die beschädigten Textsequenzen mit möglichst hoher Genauigkeit zu transkribieren. Dieser Ansatz nutzt die Eigenschaften von ATR-Verfahren: Sie arbeiten nicht mit einer reinen Zeichenerkennung, bei welcher sie Zeichen in der Vorlage isolieren und dann versuchen, eine Entsprechung zu finden und digital auszugeben. Vielmehr zerlegt das Texterkennungsmodell die einzelnen Textzeilen in kleinere Pixel-Cluster und bestimmt anhand von Mustern in der Pixelanordnung die wahrscheinlichste Buchstabenentsprechung (vgl. Graves/Schmidhuber 2008, S. 553-553). Wir versuchen, mithilfe eines nahezu allographetischen (vgl. Robinson/Solopova 1993, S. 22 sowie Stutzmann 2010, S. 250) Trainingsdatensatzes dem Modell beizubringen, auch beschädigte und schwer entzifferbare Zeichen der Handschrift zu transkribieren, welche anschließend in einer menschlichen Kontrolle überprüft und ggf. korrigiert werden. In unserem Vortrag möchten wir insbesondere zwei Aspekte näher beleuchten. Da es sich um einen experimentellen Ansatz handelt, möchten wir erstens Nutzen und Grenzen des von uns gewählten Ansatzes ausloten. Dieser ermöglicht es, mit verhältnismäßig geringem (insbesondere auch finanziellem) Aufwand beschädigte Textstellen lesbar zu machen und ist nicht auf Approximationen wie im Falle einer KI-gestützten visuellen Rekonstruktion der Vorlage durch ein Diffusionsmodell angewiesen. Gleichzeitig ist jedoch zu klären, bis zu welchem Beschädigungsgrad sich so für das menschliche Auge nicht mehr zu entziffernder Text wiederherstellen lässt. Ein zweiter Punkt widmet sich der Evaluation: Gängige Metriken für ATR-Verfahren wie die Character Error Rate (CER) messen die Genauigkeit eines Transkriptionsmodells, indem sie den Anteil von Transkriptionsfehlern an den insgesamt zu transkribierenden Zeichen messen. Als Fehler gelten dabei Hinzufügungen, Auslassungen und Veränderungen, welche das Modell im Vergleich zu einer menschlich abgesicherten Idealtranskription (einer ‚Ground Truth') ausgibt (vgl. Schonhardt et al. 2026, S. 101-102). Im Falle des Wasserschadens ist es jedoch nicht immer möglich, eine vollständig verlässliche Ground Truth bereitzustellen, da einzelne Zeichen ggf. für das menschliche Auge nicht erkennbar sind. Somit stellt sich die Frage, wie mit Ausgaben des Texterkennungsmodell umzugehen ist, welche menschlich nicht oder nur durch Rückschlüsse überprüfbar sind, die nach einer strikten Auslegung der CER als Evaluationsmetrik jedoch als Hinzufügung und somit als Fehler gelten müssten. Wir möchten daher diskutieren, wie sich Texterkennungsmodelle für unseren Anwendungsfall dennoch sinnvoll evaluieren lassen.

Literatur

- DHInfra.at. „Reading the unreadable: multispectral imaging of medieval Serbian charters". 2026. DHInfra.at*. https://www.dhinfra.at/2026-05-13-misha-cetinje-charters/ (Zugriff: 23.7.2026).
- Graves, Alex und Jürgen Schmidhuber. 2008. „Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks". Advances in Neural Information Processing Systems 21 - NIPS 2008, Lecture Notes in Computer Science*, S. 549–558. https://doi.org/10.1007/978-3-540-74690-4_56.
- Kiessling, Benjamin. 2019. „Kraken - a Universal Text Recognizer for the Humanities". Paper presented at Digital Humanities Conference 2019 (DH2019). https://doi.org/10.34894/Z9G2EX (Zugriff: 15.1.2025).

- Robinson, Peter und Elizabeth Solopova. 1993. „Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue". The Canterbury Project Occasional Papers I, S. 19–52. https://doi.org/10.5281/zenodo.11954056.
- Schonhardt, Michael et al. 2026. Automated Text Recognition. Theory, Platforms, Best-Practices. Bielefeld University Press. https://doi.org/10.64136/nrlj5339.
- Stokes, Peter et al. 2021. „The eScriptorium VRE for Manuscript Cultures. Ancient Manuscripts and Virtual Research Environments". Classics@ 18 (Ancient Manuscripts and Virtual Research Environments). https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/ (Zugriff 13.1.2025).
- Stutzmann, Dominique. 2010. „Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin?" In Kodikologie und Paläographie in digitalen Zeitalter 2*, herausgegeben von Franz Fischer, Christiane Fritze, und Georg Vogeler. Norderstedt.
- Yang, Zhenhua et al. 2025. „Predicting the Original Appearance of Damaged Historical Documents". 39th AAAI Conference on Artificial Intelligence*, Online-Vorab-Publikation. https://doi.org/10.48550/arXiv.2412.11634.

Digitalisierung zugänglich machen: Der Bildgebungs-Workflow für die hellenistischen Papyri der Bancroft Library

Moritz Eßer, Universität zu Köln; Börger Kiss, Universität zu Köln; Claes Neuefeind, Universität zu Köln; Marcel Schaeben, Universität zu Köln; Prianka Schmidt, Universität zu Köln

Lightning Talk

Das Center for the Tebtunis Papyri (CTP) an der Bancroft Library der University of California, Berkeley, hält mit rund 20.000 Objekten den bei weitem größten Bestand an hellenistischen Papyri, der zudem zu über 90 Prozent unpubliziert ist. Alle Stücke stammen aus derselben Grabungskampagne von 1899/1900 und wurden als Kartonage- und Füllmaterial für Krokodil- und Menschenmumien wiederverwendet. Fragmente, die heute getrennt vorliegen, könnten deshalb ursprünglich zu denselben Schriftstücken gehört haben. Diese wieder zusammenzuführen ist das Ziel eines auf 15 Jahre angelegten Akademievorhabens des Instituts für Altertumskunde der Universität zu Köln und der Forschungsstelle Papyrologie der Universität Münster; die digitalen Arbeitsbereiche verantwortet das Cologne Center for eHumanities (CCeH), Kooperationspartner vor Ort in Berkeley ist das CTP. Die Zusammenführung der Papyrusfragmente soll mithilfe von KI-gestützten Methoden des Fragment-Matching erfolgen. Eine wesentliche Voraussetzung für den Einsatz entsprechender Verfahren ist eine besonders einheitliche Datengrundlage. Wie diese erstellt wird und welche Überlegungen hinter der Erfassung stehen, behandelt dieser Beitrag.

Veränderte Anforderungen an die Aufnahme In Papyrussammlungen ist Reprofotografie etabliert, ihr Maßstab dabei in der Regel die bestmögliche Lesbarkeit für das menschliche Auge. Sobald die Aufnahmen als Eingangsdaten für algorithmische Vergleiche dienen sollen, verschiebt sich dieser Maßstab jedoch. Wie stark die Verfahren an der Bildqualität hängen, zeigen die einschlägigen Ansätze, etwa „siamese networks" wie Papy-S-Net (Pirrone et al. 2019), dessen Erweiterung durch self-supervised deep metric learning (Pirrone et al. 2021), graph-based reconstruction (Ostertag/Beurton-Aimar 2021) und transformer-based similarity measurement (Vu/Beurton-Aimar 2024) arbeiten alle direkt auf den Bilddaten. Besonders aufschlussreich sind die Ansätze auf Basis der Faserstruktur: Abitbol et al. (2021) führen Fragmente über den Verlauf der Papyrusfasern zusammen, und auch Carrara et al. (2026) gewinnen ihre Bildmerkmale vor allem aus der Kontinuität des Faserverlaufs.

Beide sind damit unmittelbar abhängig von einer hohen Auflösung und einer gleichmäßigen Ausleuchtung der Aufnahmen. Dass inzwischen oftmals auch synthetische Trainingsdaten erzeugt werden (Swindall et al. 2022), verweist auf die Knappheit belastbaren Bildmaterials. Für den Vergleich zählt daher weniger die Qualität der einzelnen Aufnahme als vielmehr die Konsistenz: Reproduzierbarkeit über personelle Wechsel und Gerätewartung hinweg ist damit eine zentrale technische Anforderung.

Vorbereitung und Erprobung Ein unmittelbares Vorbild lag nicht vor. Das nächstverwandte Vorhaben, das italienische JoinPap-Projekt, verfolgt ebenfalls die Zusammenführung fragmentierter Papyri und wendet sein Matching-Verfahren auf bereits digitalisierte Bestände zweier Sammlungen an (Carrara et al.2026; JoinPap 2026). In diesem Projekt dagegen bestimmte die spätere maschinelle Auswertung den Aufnahmeaufbau von Beginn an. Die konzeptionelle Vorbereitung der Digitalisierung, die vollständig vor Ort in Berkeley durchgeführt werden muss, dauerte mehrere Monate und musste fotografische wie konservatorische Aspekte gleichermaßen berücksichtigen. Mehrere Testreihen in der Kölner Papyrussammlung dienten dazu, den Aufbau unter kontrollierten Bedingungen zu erproben, bevor er vor Ort unter Zeitdruck hätte geklärt werden müssen. Eine besondere Herausforderung sind zudem die unterschiedlichen Aufbewahrungszustände: Fragmente liegen lose, zwischen Vinylite oder zwischen Glas vor, was jeweils eigene Reflexionsprobleme erzeugt und differenzierte Beleuchtungslösungen erfordert. Hinzu kommt eine erhebliche Spannweite der Objektformate, die feste Kamerapositionen ebenso betrifft wie die Handhabung der teils äußerst fragilen Stücke. Im Juni 2026 wurden vor Ort in Berkeley zwei Aufnahmestationen für Aufnahmen unter sichtbarem Licht und für Infrarot eingerichtet. Gemeinsam mit dem parallel dazu eingesetzten Flachbettscanner, entstehen so von jedem Fragment mehrere Aufnahmen, um bei der späteren Auswertung möglichst viele Bildeigenschaften einbeziehen zu können. Die Kamerastation unter sichtbarem Licht liefert eine deutlich höhere Auflösung, eine zweite im Infrarot macht verblasste oder überlagerte Tinte lesbar, und der Flachbettscanner gibt die Oberflächenstruktur besser wieder. Welche dieser Eigenschaften die besten Trainingsdaten liefern, ist noch offen, weswegen zunächst beide Verfahren – Fotografie und Flachbettscan – parallel eingesetzt werden. Ob das dauerhaft nötig ist, bleibt abzuwägen, auch mit Blick auf Zeit und Aufwand, da die Kameraaufnahmen allein deutlich schneller wäre. Statt einer Standardanwendung kommt mit „Crocodile Capture“ eine eigens am CCeH entwickelte Software zum Einsatz, die Belichtung, Farbprofil und Dateibenennung an beiden Stationen nach festen Vorgaben regelt. Ob die Aufnahmeparameter eingehalten werden, hängt damit nicht allein von der Sorgfalt der bedienenden Person ab, sondern ist im System selbst verankert. Zusätzlich sind Qualitätssicherungsmaßnahmen, wie z. B. eine automatisierte Kontrolle der Schärfe nach jeder Aufnahme, in die Software integriert.

Betrieb, Dokumentation und Ausblick Seit Juni 2026 wird die Digitalisierung vor Ort in Berkeley durch wechselnde Mitarbeitende erfolgreich durchgeführt, die technische Betreuung erfolgt von Köln aus. Wegen der Zeitverschiebung wurde hierfür eine ausführliche, auch ohne technischen Hintergrund verständliche Handreichung zu allen Arbeitsschritten erstellt, ergänzt durch Foto- und Videomaterial, die fortlaufend weiter ausgearbeitet wird. Mittelfristig sollen die Digitalisate nicht allein als Publikationsgrundlage, sondern als Trainingskorpus dienen, zunächst für bereits erprobte KI-Verfahren wie die oben genannten, perspektivisch dann auch für eigens trainierte, domänenspezifische Modelle. Weil sich das Vorhaben bewusst nicht auf eine einzelne technische Lösung festlegt, muss das Bildmaterial auch für heute noch nicht absehbare Verfahren geeignet sein. Das rechtfertigt den Aufwand für Kalibrierung, Standardisierung und Dokumentation.

Literatur

- Abitbol, Roy, Ilan Shimshoni, und Jonathan Ben-Dov. 2021. „Machine Learning Based Assembly of Fragments of Ancient Papyrus“. Journal on Computing and Cultural Heritage 14 (3): Art. 33, 1–21. https://doi.org/10.1145/3460961.
- Carrara, Fabio, Massimiliano Corsini, Fabrizio Falchi, und Nicola Messina. 2026. „JoinPap: Learning-Based Matching for the Reconstruction of Fragmentary Papyri“. In Image Analysis and Processing – ICIAP 2025 Workshops, Lecture Notes in Computer Science 16170, 296–306. Cham: Springer. https://doi.org/10.1007/978-3-032-11381-8_25.
- JoinPap. o. J. „Reconstructing Fragmentary Papyri through Human-Machine Interaction: Case Studies from Two Italian Collections“. PRIN PNRR 2022, Università degli Studi di Firenze, Istituto Papirologico „G. Vitelli“. Zugriff: 23. Juli 2026. https://www.joinpap.unifi.it/.
- Ostertag, Cecilia, und Marie Beurton-Aimar. 2021. „Using Graph Neural Networks to Reconstruct Ancient Documents“. In Pattern Recognition. ICPR International Workshops and Challenges, Lecture Notes in Computer Science 12667, 39–53. Cham: Springer. https://doi.org/10.1007/978-3-030-68787-8_3.
- Pirrone, Antoine, Marie Beurton-Aimar, und Nicholas Journet. 2019. „Papy-S-Net: A Siamese Network to Match Papyrus Fragments“. In Proceedings of the 5th International Workshop on Historical Document Imaging and Processing (HIP '19), 78–83. New York: ACM. https://doi.org/10.1145/3352631.3352646.
- Pirrone, Antoine, Marie Beurton-Aimar, und Nicholas Journet. 2021. „Self-Supervised Deep Metric Learning for Ancient Papyrus Fragments Retrieval“. International Journal on Document Analysis and Recognition (IJDAR) 24 (3): 219–234. https://doi.org/10.1007/s10032-021-00369-1.
- Swindall, Matthew I., Timothy Player, Ben Keener, Alex C. Williams, James H. Brusuelas, Federica Nicolardi, Marzia D'Angelo, Claudio Vergara, Michael McOsker, und John F. Wallin. 2022. „Dataset Augmentation in Papyrology with Generative Models: A Study of Synthetic Ancient Greek Character Images“. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22), 4973–4979. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2022/689.
- Vu, Manh-Tu, und Marie Beurton-Aimar. 2024. „ViT-ED: Transformer Network for Image Similarity Measurement“. In Document Analysis and Recognition – ICDAR 2024, Lecture Notes in Computer Science 14807, 306–323. Cham: Springer. https://doi.org/10.1007/978-3-031-70546-5_18.

Offen ist nicht gleich zugänglich. Das Corpus der barocken Deckenmalerei zwischen Retrodigitalisierung, Wissensvernetzung und KI

Theresa Baumann, LMU München / Institut für Kunstgeschichte; Maximilian Kristen, LMU München / Institut für Kunstgeschichte

Lightning Talk

Digitale Forschungsprojekte versprechen einen weitreichenden Zugang zu wissenschaftlichem Wissen. Doch die digitale Verfügbarkeit eines Forschungsgegenstandes allein bedeutet noch nicht, dass dieser tatsächlich zugänglich ist. Gerade langfristig angelegte geisteswissenschaftliche Vorhaben stehen vor der Herausforderung, historisch gewachsene Wissensbestände nicht nur zu digitalisieren, sondern sie zugleich auffindbar, verständlich, nachnutzbar und für unterschiedliche Formen der Nutzung anschlussfähig zu machen. Der Beitrag nimmt diese Problematik am Beispiel des Corpus der barocken Deckenmalerei in Deutschland (CbDD) in den Blick und fragt, wie sich Zugang zu einem kunsthistorischen Forschungsbestand gestalten lässt, der ursprünglich für das Medium Buch konzipiert wurde und heute sowohl für Menschen als auch zunehmend für Maschinen erschlossen werden soll. Das 2015 begonnene CbDD knüpft an das zwischen 1976 und 2010 publizierte Corpus der barocken Deckenmalerei in Deutschland an. Die 17 Bände des sogenannten alten Corpus mit

insgesamt 6.457 Seiten dokumentieren einen umfangreichen Bestand barocker Deckenmalerei und verbinden kunsthistorische Texte mit Fotografien, Quellen, Literaturangaben und weiteren objektspezifischen Informationen. Die gedruckte Publikation ermöglichte einen systematischen Zugang zu diesem Material, ihre Struktur war jedoch notwendigerweise an die Logik des Buches gebunden. Informationen sind dort regional und topographisch organisiert, innerhalb längerer Texte verteilt und nur über Register und Inhaltsverzeichnisse miteinander verknüpft. Für digitale Such- und Forschungsfragen bleiben zahlreiche darin enthaltene Informationen daher zunächst verborgen. Die gegenwärtige Retrodigitalisierung des alten Corpus zeigt exemplarisch, dass zwischen Digitalisierung und Zugänglichkeit mehrere Transformationsschritte liegen. Sämtliche Bände wurden digitalisiert; Texte und der weit überwiegende Teil der Abbildungen konnten automatisiert extrahiert werden. Die Texterkennung wird dabei mit Ergebnissen aus Transkribus und weiteren Verfahren kombiniert. Doch ein durchsuchbares OCR-Ergebnis bildet lediglich eine erste Ebene des Zugangs. Damit aus einem digitalisierten Buch ein strukturierter Forschungsbestand wird, müssen typographische, semantische und inhaltliche Strukturen erkannt und in ein digitales Datenmodell übersetzt werden. Überschriften, Fließtexte, Fußnoten, Bildunterschriften oder Hervorhebungen müssen ebenso identifiziert werden wie Datierungen, Maße, Künstlernamen, Räume, ikonographische Themen oder Literaturangaben. Noch grundlegender ist die Transformation der ursprünglich regional und fortlaufend strukturierten Buchtexte in die objekt- und artikelbasierte Logik des digitalen Corpus. Der Vorgang macht deutlich, dass „Zugang“ kein binärer Zustand ist. Eine gescannte Buchseite ist digital verfügbar, aber nur eingeschränkt recherchierbar. Ein OCR-Text ist durchsuchbar, seine Inhalte sind jedoch noch nicht zwingend strukturiert. Erst die Identifikation und Modellierung einzelner Entitäten und Beziehungen ermöglicht es beispielsweise, Werke eines Künstlers, Darstellungen eines bestimmten ikonographischen Themas oder Ausstattungen einer bestimmten Zeit miteinander in Beziehung zu setzen. Der Übergang vom Dokument zum Datenbestand verändert damit nicht nur die technische Zugänglichkeit, sondern auch die epistemische Struktur des Materials. Eine zentrale Rolle spielen hierbei Normdaten und kontrollierte Vokabulare. Personen, Orte, Bauwerke, ikonographische Inhalte und weitere Entitäten werden nicht lediglich als Zeichenketten behandelt, sondern soweit möglich eindeutig identifiziert und miteinander verknüpft. Für die ikonographische Erschließung kommt unter anderem Iconclass zum Einsatz. Solche Verknüpfungen erhöhen die Auffindbarkeit einzelner Objekte, ermöglichen vergleichende Abfragen und schaffen zugleich Voraussetzungen für Linked-Data-Szenarien und eine über das einzelne Projekt hinausgehende Nachnutzung. Zugang bedeutet in diesem Zusammenhang also auch Interoperabilität: Informationen müssen in einer Form vorliegen, in der sie nicht ausschließlich innerhalb einer einzelnen Website verständlich und nutzbar bleiben. Gleichzeitig darf Zugänglichkeit nicht auf Maschinenlesbarkeit reduziert werden. Für unterschiedliche Nutzer sind unterschiedliche Formen der Erschließung notwendig. Ein zentrales Element des digitalen CbDD bilden deshalb wissenschaftliche Kurztexte, die komplexe Informationen zu einem Bauwerk oder einer Ausstattung auf knappem Raum zusammenführen und so einen ersten Zugang zum Gegenstand ermöglichen. Hochauflösende Fotografien erschließen Deckenmalereien, die aufgrund ihrer Lage im Raum häufig nur schwer aus der Nähe betrachtet werden können. Dreidimensionale Modelle und Anwendungen mit Virtual-Reality-Technologie gehen noch einen Schritt weiter: Sie lösen die Werke teilweise aus den Beschränkungen ihrer physischen Sichtbarkeit und ermöglichen eine räumliche und visuelle Annäherung, die am realen Objekt vielfach nicht möglich wäre. Vermittlung, wissenschaftliche Dokumentation und digitale Forschung greifen hier ineinander. Gerade dieses Nebeneinander verschiedener Zugänge verdeutlicht jedoch auch deren Grenzen. Wissenschaftliche Vollständigkeit, intuitive Nutzbarkeit und maschinelle Strukturierbarkeit folgen nicht immer denselben Anforderungen. Eine Forschungsdatenbank muss komplexe Sachverhalte präzise abbilden können; eine öffentlich zugängliche Website darf ihre Nutzer

gleichzeitig nicht mit dieser Komplexität überfordern. Hochauflösende Bilder können für Forschung und Vermittlung zentral sein, ihre Veröffentlichung wird jedoch durch Bildrechte und Nutzungslizenzen begrenzt. Historische Begriffe und etablierte kunsthistorische Klassifikationen erleichtern fachwissenschaftliche Recherchen, können aber für ein nicht spezialisiertes Publikum erklärungsbedürftig sein. Zugänglichkeit entsteht deshalb nicht allein durch technische Öffnung, sondern durch Entscheidungen darüber, welche Informationen auf welche Weise für wen sichtbar und nutzbar gemacht werden. Mit dem zunehmenden Einsatz von Künstlicher Intelligenz erhält diese Frage eine weitere Dimension. Im Rahmen der Retrodigitalisierung können KI-basierte Verfahren dazu beitragen, OCR-Ergebnisse zu verbessern, Textstrukturen zu erkennen, Metadaten aus umfangreichen Fließtexten zu extrahieren oder große Bildbestände zu erschließen. Dadurch werden Informationen zugänglich, deren manuelle Aufbereitung mit erheblichem personellem Aufwand verbunden wäre. Zugleich birgt der Einsatz generativer Systeme gerade bei wissenschaftlichen Quellen erhebliche Risiken. Sprachmodelle können Texte normalisieren und offensichtliche Fehler erkennen, sie können jedoch ebenso historische Schreibweisen vereinheitlichen, Unsicherheiten verdecken oder vermeintlich plausible Informationen erzeugen, die sich nicht aus der Quelle ableiten lassen. Die Verbesserung des Zugangs darf deshalb nicht auf Kosten der Nachvollziehbarkeit gehen. Für wissenschaftliche Anwendungen müssen automatisierte Verarbeitungsschritte dokumentierbar bleiben und Ergebnisse auf ihre Ausgangsdaten zurückführbar sein. Darüber hinaus verändern Large Language Models selbst die Bedingungen des Zugangs. Digitale Forschungsressourcen werden künftig möglicherweise nicht mehr ausschließlich über ihre eigenen Benutzeroberflächen aufgerufen. Nutzer könnten ihre Fragen zunehmend an Sprachmodelle oder KI-Agenten richten, die Informationen aus unterschiedlichen Datenquellen zusammenführen. Damit tritt neben das klassische menschliche Publikum eine neue Gruppe von Nutzern: Maschinen, die Daten auffinden, interpretieren und miteinander verknüpfen können müssen. Für digitale geisteswissenschaftliche Projekte stellt sich deshalb die Frage, ob ihre Ressourcen lediglich über eine visuelle Website zugänglich oder zugleich so strukturiert sind, dass ihre Inhalte auch außerhalb dieser Oberfläche eindeutig adressiert und verarbeitet werden können. Das CbDD bietet für diese Entwicklung ein besonders anschauliches Beispiel. Seine Aufgabe besteht nicht lediglich darin, eine gedruckte Publikationsreihe online verfügbar zu machen. Vielmehr wird ein über Jahrzehnte entstandener wissenschaftlicher Wissensbestand in eine neue Informationsarchitektur überführt. Dabei treffen historische Texte und Fotografien auf strukturierte Daten, Normdaten, semantische Verknüpfungen, 3D-Modelle und zunehmend KI-basierte Methoden. Die Retrodigitalisierung wird so zu einem Labor für unterschiedliche Formen der Zugänglichkeit. Der Beitrag möchte daher dafür plädieren, Offenheit und Zugang in den Digital Humanities stärker voneinander zu unterscheiden. Offenheit bezeichnet zunächst die Möglichkeit, auf eine Ressource zugreifen zu können. Zugänglichkeit entsteht dagegen erst durch ein Zusammenspiel von technischer Verfügbarkeit, semantischer Strukturierung, wissenschaftlicher Nachvollziehbarkeit, verständlicher Vermittlung und nachhaltiger Anschlussfähigkeit. Dabei existiert nicht der eine Zugang für alle. Forschende, interessierte Öffentlichkeit, digitale Infrastrukturen und KI-Systeme stellen unterschiedliche Anforderungen an dieselben Daten. Am Beispiel des CbDD soll diskutiert werden, wie diese unterschiedlichen Anforderungen bereits bei der Modellierung und Erschließung digitaler Forschungsressourcen berücksichtigt werden können. Die entscheidende Frage lautet damit weniger, ob ein digitales Angebot „offen für alle“ ist, sondern vielmehr: Für wen ist es auf welche Weise offen – und welche Formen von Zugang müssen wir heute gestalten, damit geisteswissenschaftliches Wissen auch morgen noch gefunden, verstanden und weiterverwendet werden kann?

Literatur

- Anthropic. 2025. „Equipping Agents for the Real World with Agent Skills.“ October 16, 2025. https://www.anthropic.com/engineering/equipping-agents-for-the-real-world-with-agent-skills.

- Bauer, Hermann, Bernhard Rupprecht, Frank Büttner, et al., eds. 1976–2010. Corpus der barocken Deckenmalerei in Deutschland. Freistaat Bayern, Regierungsbezirk Oberbayern. 15 vols. München: Hirmer.
- Boros, Emanuela, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. „Post-Correction of Historical Text Transcripts with Large Language Models: An Exploratory Study.“ In Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), 133–159. https://doi.org/10.18653/v1/2024.latechclfl-1.14.
- Brandhorst, Hans, and Etienne Posthumus. 2017. „Iconclass: A Key to Collaboration in the Digital Humanities.“ In The Routledge Companion to Medieval Iconography, edited by Colum Hourihane, 201–218. London/New York: Routledge. https://doi.org/10.4324/9781315298375-19.
- Cordell, Ryan. 2017. „‘Q i-jtb the Raven’: Taking Dirty OCR Seriously.“ Book History 20: 188–225. https://doi.org/10.1353/bh.2017.0006.
- Dişli, Meltem, et al. 2025. „Open Data Practices of Art Museums in Wikidata: A Compliance Assessment.“ Journal of Open Humanities Data 11: 71. https://doi.org/10.5334/johd.438.
- Ji, Ziwei, et al. 2023. „Survey of Hallucination in Natural Language Generation.“ ACM Computing Surveys 55 (12): 1–38. https://doi.org/10.1145/3571730.
- Kahle, Philip, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. „Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents.“ In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) 4: 19–24. https://doi.org/10.1109/ICDAR.2017.307.
- Kristen, Maximilian. 2025. „WikiFAIR/CbDD.“ Meta-Wiki. https://meta.wikimedia.org/w/index.php?title=WikiFAIR/CbDD&oldid=28073319.
- Kristen, Maximilian. 2026a. Arthistorical-Semantic-Data-Extraction. GitHub. https://github.com/kristbaum/arthistorical-semantic-data-extraction.
- Kristen, Maximilian. 2026b. Como. GitLab. https://gitlab.wikimedia.org/kristbaum/como.
- Kristen, Maximilian. 2026c. Deckenmalerei-Lab Skill. GitHub. https://github.com/kristbaum/deckenmalerei-lab-skill.
- Lindemann, David, et al. 2025. „MediaWiki-Based Tools and Services in Digital Humanities Workflows.“ In Book of Abstracts, DARIAH AE 2025 | The Past, edited by Kim Ferguson, 18. Zenodo. https://doi.org/10.5281/zenodo.16411471.
- Poljak Bilić, Ljiljana, and Kristina Posavec. 2024. „FAIRness of Research Data in the European Humanities Landscape.“ Publications 12 (1): 6. https://doi.org/10.3390/publications12010006.
- Soria Parra, David, and Justin Spahr-Summers. 2024. „Introducing the Model Context Protocol.“ November 25, 2024. https://www.anthropic.com/news/model-context-protocol.
- Wilkinson, Mark D., et al. 2016. „The FAIR Guiding Principles for Scientific Data Management and Stewardship.“ Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18.

Digital anschlussfähig werden: Datenöffnung bei einer Printedition am Beispiel der Edition der Gottschedkorrespondenz

Ute Tischer, Sächsische Akademie der Wissenschaften zu Leipzig

Lightning Talk

Die Edition der Gottschedkorrespondenz an der Sächsischen Akademie der Wissenschaft zu Leipzig (2000-2032) ist als reine Print-Edition konzipiert. Dem entspricht ein Workflow, der sich auf das gedruckte Ergebnis konzentriert und alle nicht publizierten Zwischenschritte und -ergebnisse als temporäre „Arbeitsdaten“ betrachtet. Viele dieser „Arbeitsdaten“ sind analog, d.h. sie liegen in Form von Papierkopien oder auch Microfilmen vor, ebenso wie die zu edierenden Objekte, der Briefnachlass des Ehepaares Gottsched in der Universitätsbibliothek Leipzig. Zu Projektbeginn im Jahr 2000 standen Forschungsliteratur, Quellen und Archivalien größtenteils noch nicht digital zur Verfügung, viele Recherchearbeiten erfolgten vor Ort und ein wichtiger Teil der frühen Projektarbeit war der Aufbau

eines Handapparats in der Arbeitsstelle. 26 Jahre nach Projektbeginn und in einer nunmehr von digitalen Arbeitsformen und Ressourcen geprägten Forschungswelt stellt sich jedoch auch für das noch mehrere Jahre laufende Print-Editionsprojekt die Frage nach dem Umgang mit digitalen Daten. Solche nämlich werden, trotz des Printparadigmas, bereits seit dem Beginn der Editionsarbeit als „Arbeitsdaten“ produziert oder sind im Verlauf der Projektgeschichte hinzugekommen. Dazu zählen beispielsweise mehrere Projektdatenbanken, in denen Metadaten gesammelt wurden (Briefe, Personen, Literatur), die inzwischen vom Verlag als PDF- oder ePUB-Faksimiles angebotenen digitalen Versionen der gedruckten Bände, aber auch die Textversionen der edierten Briefe und Briefbände, die nun von der UBL bereitgestellten Digitalisate der zu edierenden Dokumente und große Mengen an digitaler Primär- und Sekundärliteratur in den verschiedensten Formaten. Mit dem Voranschreiten der Print-Edition sind auf diese Weise digitale Datensammlungen mit hohem Wert für die editorische Arbeit entstanden, die bisher aber fast ausschließlich nur projektintern genutzt werden. Unter dem Aspekt der Offenheit, Nachhaltigkeit und Langzeitverfügbarkeit von Forschungsergebnissen im digitalen Zeitalter wird inzwischen an alle Forschungsformen die wissenschaftspolitische Forderung gestellt, produzierte und gesammelte digitale Daten zu sichern und für die weitere Forschung zur Verfügung zu stellen. Auch das Projekt der Gottschededition erhielt im Zusammenhang mit der letzten externen Evaluierung 2022 die Aufforderung, die Edition „besser auf digitale Anschlussforschung einzustellen“. Das Thema des Lightning Talks wird sein, wie das Editionsteam diese Forderung umzusetzen versucht und welche Probleme sich dabei stellen. Nach einer kurzen Einführung in die editorischen und forschungspragmatischen Kontextbedingungen und einer Skizze der Ausgangssituation wird ein dreistufiger Zielekatalog vorgestellt, in dem formuliert ist, welche Maßnahmen zur Datenöffnung kurzfristig, mittelfristig und langfristig bis oder nach Projektende umgesetzt werden sollen. Nicht zu den Zielen gehörte dabei aus forschungspragmatischen Gründen ausdrücklich die Umstellung der gedruckten auf eine digitale Edition. Auf Grundlage des Zielekatalogs entstand ein Arbeits- und Zeitplan für die Umsetzung während und zusätzlich zur laufenden Projektarbeit an der Printedition. Die wichtigste Erkenntnis bei der Erarbeitung und Erprobung dieser Konzepte war, dass Offenheit und Zugänglichkeit entsprechende Datenkuration voraussetzen, die im Projekt zuvor zu leisten und umzusetzen ist. Ein solches Forschungsdatenmanagement ist zeitintensiv und setzt Kompetenzen und Infrastrukturen voraus, die in einem Printprojekt nicht selbstverständlich vorhanden und nicht Teil der Projektförderung sind. Im letzten Teil des Lightning Talks werden exemplarisch die Maßnahmen beschrieben, mit denen das Projekt Briefmetadaten Open Access digital zur Verfügung stellt. Mit den geschilderten Problemen dürfte die Gottschededition beispielhaft für viele andere aktuelle Langzeit-Printeditionen im Akademienprogramm stehen, die über ihr eigentliches Produkt hinaus das digitale Paradigma bedienen müssen, um der modernen, digital basierten geisteswissenschaftlichen Forschung rezipiert zu werden. Der Lighning Talk möchte Impulse geben und zum Erfahrungsaustausch anregen.

Literatur

- Johann Christoph und Luise Adelgunde Victorie Gottsched. 2026. Briefwechsel. Historisch-kritische Ausgabe. Band 21: April 1753 − August 1753. Berlin, Boston: de Gruyter 2026.
- Brehm, Elke und Janna Neumann. 2018. „Anforderungen an Open-Access-Publikationen von Forschungsdaten – Empfehlungen für einen offenen Umgang mit Forschungsdaten.“ o.bib 5 (3): 1–16. DOI: 10.5282/o-bib/2018H3S1-16.
- Lemke, Karoline, Sandra König, Harald Lordick und Gerrit Brüning. (2025) 2025. „Empfehlung zur Erstellung, Bearbeitung und Publikation FAIRer Forschungsdaten in der Datendomäne Editionen. A Living Handbook über das How-to der Bewertung und Gewährleistung FAIRer, qualitativ hochwertiger editionswissenschaftlichen Forschungsdaten.“ https://textplus.pages.gwdg.de/textplus-editions/guidelines_sde/.

- Münkner, Jörn, Maximilian Görmar und Joëlle Weis, Hrsg. 2024. Sammlung und Netz. Theoretische und praxeologische Implikationen. Kulturen des Sammelns 6. Göttingen: Wallstein Verlag.
- Rat für Informationsinfrastrukturen. 2021. Bestandsbezogene Forschung gestalten: zukunftsfähige Verschränkungen von „digital“ und „analog“. Ein Diskussionsimpuls zur wissenschaftlichen, wissenschaftsnahen und kulturellen Nutzbarkeit von Sammlungen. Göttingen: Rat für Informationsinfrastrukturen. https://d-nb.info/1247657612/34.
- Regeln der Sächsischen Akademie der Wissenschaften zu Leipzig zur Sicherung guter wissenschaftlicher Praxis. 2023. https://www.saw-leipzig.de/de/ueber-die-akademie/satzungen-und-ordnungen/regeln-der-saw-zur-sicherung-guter-wissenschaftlicher-praxis.pdf.
- Tischer, Ute. 2025. Digitale Workflows bei Langzeit-Printeditionen. Die Edition der Gottschedkorrespondenz. Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft 564. Berlin: Humboldt-Universität zu Berlin. https://doi.org/10.18452/35508.
- Wissenschaftsrat. 2020. Zum Wandel in den Wissenschaften durch datenintensive Forschung. Berlin. https://www.wissenschaftsrat.de/download/2020/8667-20.html.
- Wissenschaftsrat. 2020. Wissenschaftspolitische Stellungnahme zum Akademienprogramm. Berlin. https://www.wissenschaftsrat.de/download/2020/8287-20.

Einblick in die Automatisierungsworkflows in EuReD

Kevin Kuck, ULB Darmstadt; Kevin Wunsch, ULB Darmstadt

Lightning Talk

Obwohl jede digitale Edition spezifische Anforderungen hat, sind viele Arbeitsschritte identisch oder ähnlich genug für eine schematische Erfassung und (halb-)automatische Ausführung. Nach Hilary Mason liegt „the best part of programming“ im „triumph of seeing the machine do something useful“ („Automate the Boring Stuff with Python - Automate the Boring Stuff with Python“, n. d.). Das große Automatisierungs- und Standardisierungspotential im Editionsprozess hat das Projekt „The Flow“ hervorgehoben. Dem Ziel standardisierte Workflows auf existierenden Lösungen aufzubauen haben wir uns ebenso verschrieben wie „The Flow“ („The Flow“, n. d). Optimale Ausgangspunkte wären der Up- und Download unterschiedlicher Formate an gleichbleibende Orte, etwa den Netzwerkspeicher, oder die Publikationsumgebung auch die Texterkennung mittels Transkribus, oder ähnlichen Tools. Durch die Standardisierung und Automatisierung solcher Routineaufgaben können sich Forschende auf die eigentlichen inhaltlichen und Entwickler*innen auf die konzeptionellen Herausforderungen konzentrieren, während die Maschine die repetitiven Arbeitsschritte übernimmt. Dies führt zu einer effizienteren Ressourcennutzung im gesamten Editionsprozess: Durch die Teilautomatisierung von Transkription, Upload und Hosting kann etwa die Arbeit am kritischen Apparat umfangreicher werden und die digitale Forschung vereinfacht werden (cf. Terras et al. 2024, 4).

Dabei sind besonders die OCR-Erfassung der Texte mittels Transkribus, das Importieren in die Arbeitsumgebung und die abschließende Veröffentlichung, sowie Sicherung auf Gitlab und Langzeitarchivierung hervorzuheben. Diese Schritte sind für jede Edition in Abhängigkeit vom genutzten Framework, etwa Ediarum, der teiPublisher, oder wdbplus, relativ stromlinienförmig. Ähnliche oder gar gleichartige Schritte laden jede\n Entwickler\in ein, diese ‘langweiligen’ Arbeitsschritte noch gleichartiger zu machen und sie, frei nach Al Sweigart’s „Automate the Boring Stuff“, zu automatisieren. Wir gehen mit Europäische Religionsfrieden Digital („Europäische Religionsfrieden“, n. d.) diesen Schritt und hoffen, durch die Nutzung des Automatisierungspotentials langfristig den Editionsprozess effizienter zu gestalten und den Fokus verstärkt auf die spannenderen Schritte, die Bereitstellung des kritischen Apparates und die digitale Forschung, zu legen. Wo Editor\innen von Transkribus unterstützt werden sollten Entwickler\innen ihre Zeit nicht mit repetitiven Aufgaben, wie etwa dem wiederholten Down- bzw. Upload von Dateien verbringen, sondern sich der Erforschung

der Gegenstände mit digitalen Methoden widmen. Ausgehend von den Überlegungen innerhalb von EuReD entwickeln wir einen teilautomatisierten Workflow, der den gesamten Entstehungsprozess vom Digitalisat bis hin zum angereicherten Volltext einer digitalen Edition abbilden soll. Zur Orchestrierung des Workflows wird die Python-basierte Bibliothek Prefect eingesetzt, die eine ereignisgesteuerte und sequenzielle Ausführung einzelner Funktionen ermöglicht. Durch den Fokus auf Python wird eine weit verbreitete und leicht zu erlernende Sprache genutzt. Da Prefect eine selbst-gehostete Version anbietet, fallen keine weiteren Kosten an und die Daten sind zu jedem Zeitpunkt DSGVO-konform auf den eigenen Servern. Nachdem das Digitalisat im Projekt angekommen ist, wird dieses in einem ersten Schritt auf dem Netzwerkspeicher abgelegt, von wo der Scan automatisiert nach Transkribus importiert wird. Hierfür erfolgen verschiedene API-Abfragen zur Interaktion mit Transkribus, etwa für Login, Authentifizierung, Dateiübertragung, Layoutanalyse sowie Texterkennung. Sobald der Text in möglichst fehlerfreier Form vorliegt, wird der Exportprozess gestartet. Dabei wird die erzeugte TEI-Fassung automatisiert gegen die projektspezifischen Schema-Dateien validiert. Nur der validierte Text wird je nach Stand des Workflows entweder auf dem Netzwerkspeicher oder direkt in der Arbeitsumgebung abgelegt, wo weitere manuelle Bearbeitungsschritte erfolgen, wie beispielsweise weitere Korrekturen oder das Erstellen des kritischen Apparats. Fehlerhafte Texte haben eine Meldung an das Projektteam zur Folge, sodass die Fehler manuell behoben werden können. Perspektivisch soll an dieser Stelle ein halbautomatischer Schritt zur Named Entity Recognition integriert werden, der die editorische Arbeit an Texten unterstützt oder ersetzt. Die maschinell erkannten Entitäten können dann von Editor\innen überprüft, ergänzt und gegebenenfalls korrigiert werden. Im Anschluss wird der fertige Text über einen weiteren Prefect-Flow automatisch in das zugehörige Git-Repository überführt und auf dem Präsentationsserver persistent zitierbar veröffentlicht. Dabei erfolgt nicht nur eine abschließende Validierung, sondern auch eine abschließende Homogenisierung der Daten. Bei Fehlern wird der Prozess abgebrochen und dokumentiert, sodass veröffentlichte Dateien stets den Standards des Zentrums für digitale Editionen entsprechen. Zukünftig soll für jedes Projekt eine RDF-Darstellung hergestellt und über den an der ULB Darmstadt entwickelten rdfStore bereitgestellt werden (Rokit / RDF Store 2025). Grundlage der Automatisierungen und Transformationen sind dabei die Konversion und Validierung von Texten, die Sicherung und Dokumentation auf GitLab und GitHub. Dabei wird jeder Workflow-Schritt automatisch mit einem Git-Issue im internen GitLab dokumentiert, bei Änderungen aktualisiert und nach erfolgreicher Ausführung geschlossen, um Nachvollziehbarkeit und Kontrolle zu gewährleisten. Zukünftig werden zudem alle veröffentlichten Texte automatisch auf GitHub der Öffentlichkeit zur Verfügung gestellt. Für Automatisierungen sind standardisierte Projektstrukturen essentiell: Projektordner und Dateinamen müssen standardisiert sein, Validierungsskripte und Transformationen sollten innerhalb einzelner Projekte an festen Pfaden befinden und müssen lediglich an projektspezifische Anforderungen angepasst werden, um eine zuverlässige wiederholbare Ausführung der automatisierten Prozesse zu ermöglichen. Prefect liefert ein rudimentäres Frontend, mit dem die Flows und Tasks ausgeführt werden können. Mittelfristig soll ein Control Center entstehen, in dem die Handhabung der Automatisierung einfach möglich sein soll. Dies soll auch Kolleg\innen, die nicht mit den jeweiligen Projektdetails vertraut sind, eine niedrigschwellige Überwachung ermöglichen.

Literatur

- „Automate the Boring Stuff with Python.“ n.d. Accessed August 26, 2026. https://automatetheboringstuff.com/.
- „Europäische Religionsfrieden.“ n.d. Accessed August 26, 2026. https://eured.de/.
- Rokit / RDF Store. 2025. „Rokit / RDF Store.“ Accessed August 26, 2026. https://gitlab.ulb.tu-darmstadt.de/rokit/rdf-store.

- Terras, Melissa, Joe Nockels, Sarah Ames, Paul Gooding, Andy Stauder, and Günter Mühlberger. 2024. „On Automating Editions.“ Scholarly Editing Journal 41. https://doi.org/10.55520/W257A74E.
- „The Flow.“ n.d. Accessed August 26, 2026. https://www.flow-project.net/.

Session 4: Vernetzung, Knowledge Graphs & Suche

Offener Zugang trotz Komplexität – LLM-basierte Assistenzsysteme am Beispiel wissenschaftlicher Suchanwendungen

Thomas Eckart, SAW Leipzig; Uwe Kretschmer, SAW Leipzig

Vortrag

Digitale Forschungsinfrastrukturen müssen sich an den Bedürfnissen unterschiedlichster Nutzergruppen orientieren und ihre Angebote entsprechend ausgestalten. Dabei ergeben sich Unterschiede nicht nur aus den verschiedenen fachlichen Interessen, sondern ebenso aus unterschiedlichem Vorwissen, allgemeinen technischen und sprachlichen Kompetenzen, und der jeweiligen Erwartungen an die konkrete Anwendung. Wissenschaftliche Informationssysteme stehen damit vor einem grundlegenden Problem: Sie sollen einen möglichst einfachen Zugang für eine Vielzahl von Personengruppen ermöglichen, ohne die für konkrete Forschungsinteressen notwendige funktionale Komplexität zu reduzieren.

Diese Herausforderung zeigt sich auch bei verschiedenen Anwendungen des NFDI-Konsortiums Text+, das primär den Bereich sprach- und textbasierter Forschungsdaten adressiert. Ein konkretes Beispiel hierfür ist die Föderierte Inhaltssuche (Federated Content Search, kurz: FCS)(Uytvanck et al. 2017, Körner und Eckart 2025), die eine parallele, verteilte Suche über die Inhalte verschiedenster Ressourcen ermöglicht, die von diversen wissenschaftlichen Einrichtungen bereitgestellt werden. Die FCS unterstützt die Suche in unterschiedlichen Arten von Forschungsdaten und adressiert deren Unterschiede durch Ressourcentyp spezifische Darstellungsformen und formale Anfragesprachen. Zum Beispiel werden Textkorpora und wissenschaftliche Editionen mit der Anfragesprache FCS-QL durchsucht, lexikalische Ressourcen wie Wörterbücher oder lexikalisch-semantische Wortnetze mit dem CQL-Dialekt LexCQL (Körner et al. 2025). Beide Anfragesprachen ermöglichen komplexe Recherchen in den angeschlossenen Datenbeständen, setzen jedoch Kenntnisse über deren jeweilige Syntax, unterstützte Suchfelder, Operatoren usw. voraus.

In welchem Ausmaß technische Anforderungen und die konkrete Ausgestaltung von FCS-Nutzeroberflächen deren Nutzung beeinflussen, ist Gegenstand einer 2023 begonnenen Veranstaltungsreihe mit bisher sechs Hackathons und Workshops, in denen gezielt der Austausch mit verschiedenen Nutzergruppen gesucht wurde. Die dabei formulierten Anforderungen (Eckart et al. 2025) betreffen unter anderem die Leistungsfähigkeit der Anwendungen, die Ergebnispräsentation sowie die Einbindung von Suchverfahren in weiterführende Forschungsprozesse. Als übergreifendes Problem zeigte sich jedoch wiederholt, dass vorhandene Funktionen und Ausdrucksmöglichkeiten für Nutzer:innen nicht ausreichend sichtbar oder unmittelbar verständlich sind. Gleichzeitig kann insbesondere bei der erstmaligen Nutzung einer Anwendung nicht erwartet werden, dass Nutzer:innen umfangreiche technische Spezifikationen konsultieren. Zentrale Funktionen und Anfragemöglichkeiten sollten möglichst direkt über die Nutzeroberfläche erschließbar sein.

Als häufig genannte Einstiegshürde erwies sich die Formulierung valider Anfragen an das Informationssystem. Probleme zeigen sich hier auf syntaktischer Ebene (Welche Operatoren erlaubt die jeweilige Anfragesprache? Wie kombiniere ich Suchterme miteinander?), aber insbesondere auch auf semantischer Ebene (Was sind die konkreten Bedeutungsunterschiede zwischen verschiedenen Anfrageformen? Welche Annotationen bzw. Vokabulare werden verwendet?). Über die verschiedenen Workshops hinweg zeigte sich entsprechend ein Bedarf an leicht zugänglichen Hilfen zur Anfrageformulierung sowie an kontextbezogener Unterstützung direkt in der Suchoberfläche. Als mögliche Unterstützungsangebote wurden wiederholt Hilfeseiten, FAQs, Cheat Sheets, die Bereitstellung valider Beispielanfragen im UI oder visuelle Query Builder diskutiert.

Die Nutzung großer Sprachmodelle bietet eine Möglichkeit, derartige Unterstützungsangebote zu ergänzen (Mayerhofer et al. 2025). Der hier diskutierte Ansatz besteht dabei nicht darin, formale Anfragesprachen hinter einer Chatoberfläche zu verbergen. Gerade für wissenschaftliche Anwendungen besitzen formale Anfragen wichtige Eigenschaften: Sie sind validierbar und eindeutig, ihre Semantik ist nachvollziehbar und Suchvorgänge lassen sich reproduzieren. KI-basierte Unterstützung sollte deshalb eher als optionale „Co-Pilot“-Funktion innerhalb bestehender Anwendungen umgesetzt werden. Sie kann insbesondere unerfahrene Nutzer:innen bei der Formulierung komplexer Anfragen unterstützen, indem ein zunächst natürlichsprachlich formuliertes Informationsbedürfnis in eine formale Anfrage überführt wird. Diese bleibt dabei jederzeit sichtbar und editierbar. Ziel ist nicht, eine dauerhafte Abhängigkeit von der LLM-Unterstützung zu schaffen, sondern Nutzer:innen schrittweise mit den Möglichkeiten und der Logik des jeweiligen Systems vertraut zu machen. Ähnliche Ansätze existieren bereits für andere formale Anfragesprachen, beispielsweise SPARQL oder SQL (Enguix 2024; Schneider et al. 2024).

Auch in der Gegenrichtung kann eine solche Komponente unterstützend eingreifen: Komplexe formale Anfragen lassen sich in einer natürlichsprachlichen Umschreibung häufig leichter erfassen und auf Übereinstimmung mit dem konkreten Suchinteresse überprüfen. Eine automatisch erzeugte Erklärung soll damit nicht nur beim Verständnis einzelner bestehender Anfragen helfen, sondern den Lernprozess generell unterstützen.

Für die FCS wurde dieser Ansatz für die dort verwendeten Anfragesprachen, FCS-QL und LexCQL, konkret umgesetzt. Die Unterstützung arbeitet in einem Few-Shot-Setting auf Grundlage der jeweiligen formalen Spezifikation und nutzt im Hintergrund den OpenAI-API-kompatiblen LLM-Service eines wissenschaftlichen Rechenzentrums. Natürlichsprachliche Eingaben werden in eine entsprechende formale Anfrage überführt, während umgekehrt als komplex erkannte FCS-QL- oder LexCQL-Anfragen durch eine natürlichsprachliche Erklärung beschrieben werden.

Von besonderer Bedeutung ist, dass diese Funktionen nicht als zusätzliches Chat-Interface umgesetzt wurden, sondern als konkrete Komponenten in die bereits bestehende Nutzeroberfläche integriert sind. Die LLM-basierte Unterstützung wird damit unmittelbar dort angeboten, wo Nutzer:innen Anfragen formulieren, testen und überarbeiten. Damit bleibt die übliche Interaktion mit der Suchanwendung erhalten, während zusätzliche Unterstützung zur Verfügung gestellt wird: generierte Anfragen und deren formale Struktur können durch die Nutzer:innen direkt in der Anwendung nachvollzogen und bei Bedarf angepasst werden.

Wie nützlich ein solcher Assistenzansatz in der Praxis ist, hängt wesentlich davon ab, in welchem Umfang er unterschiedliche Nutzergruppen unterstützen kann. Hier spielt unter anderem die Überwindung sprachlicher Zugangshürden eine wichtige Rolle. Multilinguale Sprachmodelle ermöglichen es grundsätzlich, natürlichsprachliche Eingaben in verschiedenen Sprachen auf dieselbe formale Anfragesprache abzubilden. Die bisherigen Evaluationen zeigen jedoch, dass die Qualität sowohl

von der verwendeten Sprache als auch vom jeweiligen Modell abhängt. In einer Untersuchung zur LexCQL-Generierung wurden 280 deutsch- und englischsprachige Anfragen manuell bewertet. Insgesamt waren 81,4 Prozent der erzeugten Anfragen korrekt; bei englischen Eingaben lag dieser Wert bei 90 Prozent, bei deutschen bei 72,9 Prozent (Eckart et al. 2026a). Weiterführende Tests (Eckart et al. 2026b) mit mehreren Open-Weight-Modellen und sechs europäischen Sprachen (Deutsch, Englisch, Französisch, Italienisch, Polnisch und Spanisch) zeigen eine Tendenz zu besseren Ergebnissen bei größeren Modellen. Zugleich treten deutliche Qualitätsunterschiede zwischen den getesteten Modellen und Eingabesprachen auf; darüber hinaus erweist sich die konkrete Promptgestaltung als relevanter Einflussfaktor (zu den allgemeinen Effekten siehe auch Qin et al. 2025 bzw. Romanou et al. 2026). LLM-basierte Assistenz kann damit zusätzliche Zugänge eröffnen, setzt für den Einsatz in wissenschaftlichen Informationssystemen jedoch eine kontinuierliche, mehrsprachige und modellübergreifende Qualitätssicherung voraus.

Für wissenschaftliche Informationssysteme eröffnet dieser Ansatz ein breiteres Einsatzfeld als die automatische Generierung einzelner Anfragen. Denkbar sind auch Hinweise auf verfügbare Suchmöglichkeiten oder Unterstützung bei der Auswahl geeigneter Ressourcen. Die grundlegende Idee bleibt dabei, bestehende spezialisierte Anwendungen nicht durch generative Systeme zu ersetzen, sondern ihre Nutzung schrittweise zugänglicher zu machen.

Damit entstehen allerdings neue Zielkonflikte. In den Evaluationen zeigt sich der Zusammenhang zwischen Modellgröße und Ergebnisqualität: Größere Modelle erzielten bei der Generierung formaler Anfragen tendenziell bessere Ergebnisse und erwiesen sich insbesondere für eine zuverlässige mehrsprachige Unterstützung als leistungsfähiger. Diese Qualitätsgewinne stehen jedoch einem höheren Ressourcenbedarf gegenüber, da größere Modelle mehr Rechenleistung erfordern und sowohl Betriebskosten als auch Inferenzzeiten erhöhen können. Gerade auch Letzteres ist im Kontext der hier dargestellten Integration von hoher Bedeutung. Für Forschungsinfrastrukturen stellt sich deshalb nicht nur die Frage, wo LLM-Unterstützung generell sinnvoll ist, sondern auch, wann sie tatsächlich benötigt wird. Ressourcenschonende Architekturen können beispielsweise Assistenz nur bei erkennbar komplexen Eingaben aktivieren oder unterschiedlich große Modelle abhängig von Aufgabe und benötigter Sprachabdeckung einsetzen.

Insgesamt können LLMs den Zugang zu komplexen wissenschaftlichen Anwendungen deutlich erleichtern. Voraussetzung ist jedoch, ihre Nutzung als gezielte Assistenz zu gestalten – mit nachvollziehbaren Ergebnissen, kontrollierbarer Qualität und einem Ressourcenaufwand, der in einem akzeptablen Verhältnis zum tatsächlichen Nutzen steht.

Literatur

- Eckart, Thomas, Erik Körner, und Uwe Kretschmer. 2025. „Usability der LexFCS: Anforderungen, Problemfälle und Priorisierung.“ https://doi.org/10.5281/zenodo.17055166.
- Eckart, Thomas, Uwe Kretschmer, Erik Körner, Felix Helfer, und Hagen Beelitz. 2026. „Do you know what I’m looking for? LLM-supported Query Formulation for Complex Query Languages.“ In Book of Abstracts: EADH 2026 Conference. Krakau: Jagiellonian University. (to be published)
- Eckart, Thomas, Erik Körner, Hagen Beelitz, und Uwe Kretschmer. 2026. „Semantic Parsing and Query Explanation for the CLARIN Search Infrastructure.“ In Proceedings of the CLARIN Annual Conference 2026, CLARIN ERIC. (to be published)
- Enguix, Carlos F. 2024. „Comparison of LLM-based Natural Language to SPARQL Queries Systems.“ https://orkg.org/comparisons/R702114.
- Körner, Erik und Thomas Eckart. 2025. „Accessing Linguistic Content in Distributed Research Environments“. In Harmonising Language Data: Standards for Linguistic Resources, herausgegeben von Piotr Bański, Ulrich Heid und Laura Herzberg, [377-399]. Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783112208212.

- Körner, Erik, Thomas Eckart, Uwe Kretschmer, Axel Herold, Frank Wiegand, Frank Michaelis, Matthias Bremm, Louis Cotgrove, Thorsten Trippel, Felix Rau, Anne Klee, Daniel Werning, Dominik Blöse, und Claus Zinn. 2025. „Federated Content Search for Lexical Resources (LexFCS): Specification“. https://doi.org/10.5281/zenodo.15706299.
- Mayerhofer, Kerstin, Rob Capra, und David Elsweiler. 2025. „Blending Queries and Conversations: Understanding Tactics, Trust, Verification, and System Choice in Web Search and Chat Interactions.“ arXiv:2504.05156. https://arxiv.org/abs/2504.05156.
- Qin, Libo, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, und Philip S. Yu. 2025. „A Survey of Multilingual Large Language Models.“ Patterns 6 (1): 101118. https://doi.org/10.1016/j.patter.2024.101118.
- Romanou, Angelika, Marwa Ibrahim, Candace Ross, Chantal Shaib, Kutay Oktar, Sam J. Bell, Anaelia Ovalle, Jesse Dodge, Antoine Bosselut, Koustuv Sinha, und Adina Williams. 2026. „Brittlebench: Quantifying LLM Robustness via Prompt Sensitivity.“ arXiv:2603.13285. https://arxiv.org/abs/2603.13285.
- Schneider, Philipp, Marcel Klettner, Kristiina Jokinen, Elena Simperl, und Florian Matthes. 2024. „Evaluating Large Language Models in Semantic Parsing for Conversational Question Answering over Knowledge Graphs.“ arXiv:2401.01711. https://arxiv.org/abs/2401.01711.
- Uytvanck, Dieter van, Leif-Jöran Olsson, Oliver Schonefeld, et. al 2017. „CLARIN Federated Content Search (CLARIN-FCS) – Core 2.0“. https://office.clarin.eu/v/CE-2017-1046-FCS-Specification-v20230426.pdf.

Vernetzte Zugänglichkeit: Semantische Interoperabilität im Akademieprojekt „Steinerne Zeugen digital“

Peter Ackermann, Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte an der Universität Duisburg-Essen; Anna Martin, Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte an der Universität Duisburg-Essen

Lightning Talk

Jüdische Friedhöfe zählen zu den ältesten Zeugnissen der Sepulkralkultur in Deutschland. Das Akademieprojekt „Steinerne Zeugen digital“ (SZd), das seit 2023 am Steinheim-Institut und an der Universität Bamberg durchgeführt wird, dokumentiert diese Friedhöfe über 24 Jahre hinweg disziplinenübergreifend: Judaistik, Bauforschung und Digitale Denkmaltechnologien erfassen Inschriften, Objekteigenschaften und topografische Zusammenhänge in einem gemeinsamen, durchgängig digitalen Workflow. Der Beitrag greift die Frage der Tagung „Offen für alle?“ auf und zeigt, wie Zugänglichkeit über die reine Online-Verfügbarkeit hinaus auch die maschinenlesbare Verknüpfung von Forschungsdaten über Fachgrenzen hinweg umfassen muss.

Im Zentrum des Vortrags steht die geplante semantische Vernetzung der drei zentralen Datensysteme des Projekts: der epigraphischen Datenbank epidat, der Forschungsplattform MonArch und einer künftigen Personendatenbank. Wir stellen vor, wie Text-, Objekt- und Raumdaten künftig über gemeinsame Standards und Schnittstellen so verbunden werden, dass sie sowohl für Forschende als auch für automatisierte Nachnutzung durchsuchbar und referenzierbar sind.

Aktuell bestehen die zentralen Datensysteme des Projekts weitgehend isoliert nebeneinander: Die digitalen Editionen in epidat sind seit 2006 mit mittlerweile rund 50.000 Grabsteininschriften und 30.000 Personeneinträgen online zugänglich, die darin enthaltenen Informationen aber wenig strukturiert ausgezeichnet; in MonArch werden räumliche und objektbezogene Daten aus Bauforschung und 3D-Digitalisierung gebündelt; Archivrecherchen liegen bislang größtenteils in Excel-Listen und Word-Dokumenten vor. Die Verknüpfung zwischen diesen Ebenen erfolgt derzeit nur über wechselseitige ID-Angaben, nicht aber über ein gemeinsames semantisches Modell.

Das Projekt begegnet dieser Herausforderung mit einer mehrstufigen Interoperabilitätsstrategie. Auf der Datenmodellebene soll der epigraphische Standard EpiDoc (TEI/XML) mit einem bauforscherischen SKOS-Vokabular von rund 8.500 Begriffen sowie einer CIDOC-CRM-Anbindung verschränkt werden, sodass sich Text-, Objekt- und Raumdaten in einer gemeinsamen semantischen Struktur beschreiben lassen. Auf der Systemebene ist ein Relaunch von epidat als Epigraphik-Tool geplant, das Personen, Rollen, Ereignisangaben, Orte und Symbole direkt mit der bauforscherischen Auszeichnung in MonArch und mit einer eigenständigen Personendatenbank verlinkt, die über die bisherigen, prosopographisch noch nicht hinreichend differenzierten Personendaten in epidat hinausgeht. Auf der Infrastrukturebene sieht die MonArch-Roadmap Schnittstellen wie OAI-PMH und IIIF, eine facettierte Suche sowie eine dedizierte Linked-Open-Data-Erweiterung vor, über die perspektivisch auch eine Anbindung an GND, Wikidata und das Fachvokabular hist-arch-vocab erfolgen soll.

Anhand eines exemplarischen Grabmals lässt sich diese Vernetzung konkret zeigen: Eine Inschrift verweist auf eine bestattete Person, deren Rolle in der Gemeinde, ein Ereignis (etwa das Sterbedatum) sowie auf ein Symbol im Steindekor. Im derzeitigen System bleiben diese Informationen in getrennten, nur locker verbundenen Datensätzen verstreut. Im geplanten Modell werden Person, Rolle, Ereignis und Symbol jeweils mit der Personendatenbank beziehungsweise der bauforscherischen Auszeichnung in MonArch verknüpft, sodass sich Anfragen wie „alle Grabmale einer bestimmten Familie mit einem bestimmten Symbol in einem bestimmten Zeitraum“ erstmals systemübergreifend beantworten lassen.

Der Vortrag stellt diese Vernetzungsarchitektur als Werkstattbericht vor und ordnet sie in die Debatte um Zugänglichkeit ein: Eine konsequent semantische Verknüpfung heterogener Fachsysteme erhöht nicht nur die Auffindbarkeit für menschliche Nutzer:innen, sondern schafft auch die Grundlage dafür, geisteswissenschaftliche Daten für neue Nutzungsgruppen wie KI-Agenten überhaupt erst adressierbar zu machen. Gleichzeitig werden im Vortrag offene Probleme kritisch reflektiert: Wie granular muss ein gemeinsames Datenmodell sein, um zwischen den unterschiedlichen epistemischen Logiken von Epigraphik, Bauforschung und Prosopographie zu vermitteln, ohne einzelne Fachperspektiven zu verflachen? Und wie können wir damit umgehen, dass technische Interoperabilität allein noch keine inhaltliche oder kulturelle Zugänglichkeit garantiert?

Der Beitrag versteht sich als Diskussionsimpuls für andere Digital-Humanities-Projekte, die vor der Aufgabe stehen, historisch gewachsene, disziplinär geprägte Datenbestände nachträglich zu vernetzen, und lädt zum Austausch über Erfahrungen mit vergleichbaren Interoperabilitätsstrategien ein.

Literatur

- Arera-Rütenik, Tobias, und Thomas Kollatz. 2016. „Interdisziplinäre Perspektiven auf Grabmale und Visualisierung räumlicher Strukturen. Ergebnisse eines Projektes zu historischen jüdischen Friedhöfen.“ In Objekt und Schrift: Beiträge zur materiellen Kultur des Jüdischen* (Jüdisches Kulturerbe 1), herausgegeben von A. von Kienlin, K. Keßler, U. Knufinke, und S. M. Ross, 161–168. Braunschweig: TU Braunschweig.
- Kollatz, Thomas. 2018. „EPIDAT. Research Platform for Jewish Epigraphy.“ In Crossing Experiences in Digital Epigraphy. From Practice to Discipline*, herausgegeben von Annamaria De Santis und Irene Rossi, 231–239. Warschau. https://doi.org/10.1515/9783110607208.

API, Filter-Interface und KI-Chat: Wie historia.editionhub.org einen Wissensgraphen befragt — Ein Werkstattbericht

Hans Clausen, Freies Deutsches Hochstift Frankfurt am Main

Lightning Talk

Die Forschungsplattform historia.editionhub.org[1] aggregiert Informationen zu historischen Aktivitäten und den dazugehörigen Quellen aus digitalen Editionen. Aktivitäten und Quellen werden mittels CIDOC-CRM[2] in einem gemeinsamen Wissensgraphen repräsentiert. Neben Datum, Ort und Akteur werden auch der materielle Kontext sowie kognitive Inhalte einer wissensbezogenen Aktivität abgebildet (Renn 2012, 20–21). So lassen sich soziale, semiotische und semantische Wissensnetzwerke einzelner Akteure erfassen (Renn et al. 2016, 37–41).

Historia.editionhub.org wird im Rahmen des Projektes „PROPYLÄEN. Forschungsplattform zu Goethes Biographica“[3] entwickelt. In vier Teileditionen werden Goethes Tagebücher, etwa 15.000 überlieferte Briefe Goethes, mehr als 20.000 Briefe an Goethe sowie Zeugnisse seiner Begegnungen und Gespräche digital ediert und auf goethe-biographica.de zugänglich gemacht. Historia.editionhub.org extrahiert die in den TEI-Dokumenten[4] ausgezeichneten Aktivitäten Goethes und verknüpft sie mit ihren Quellen. Erfasst werden neben Korrespondenz, Übernachtungen und Begegnungen auch Reisen, Lektüre, Werkarbeit und Mittagessen.

Die Zeugnisse zeigen, dass Goethe nicht nur ein bedeutender Schöpfer deutschsprachiger Dichtung, sondern auch ein vielseitig tätiger Akteur in Kunst, Politik und Wissenschaft war. Um Goethes Einbindung in die gesellschaftlichen Wissensstrukturen seiner Zeit sichtbar zu machen, sollen die Aktivitäten möglichst vieler seiner Zeitgenossen und Kontaktpersonen nach den beschriebenen Modellierungskriterien zusammengeführt werden. Dafür wurde das TEI-konforme Austauschformat HASE (Historical Activity and Source Exchange)[5] entwickelt, über das sich Aktivitäten und ihre Quellen aus weiteren Editionsprojekten niederschwellig in die Plattform integrieren lassen. Dadurch soll der Community in Zukunft nicht nur ein besseres Verständnis von Goethe, sondern auch ein umfassenderes Verständnis seiner Zeit ermöglicht werden.

Für die Dissemination der zusammengeführten Informationen stehen drei Zugänge zur Verfügung:

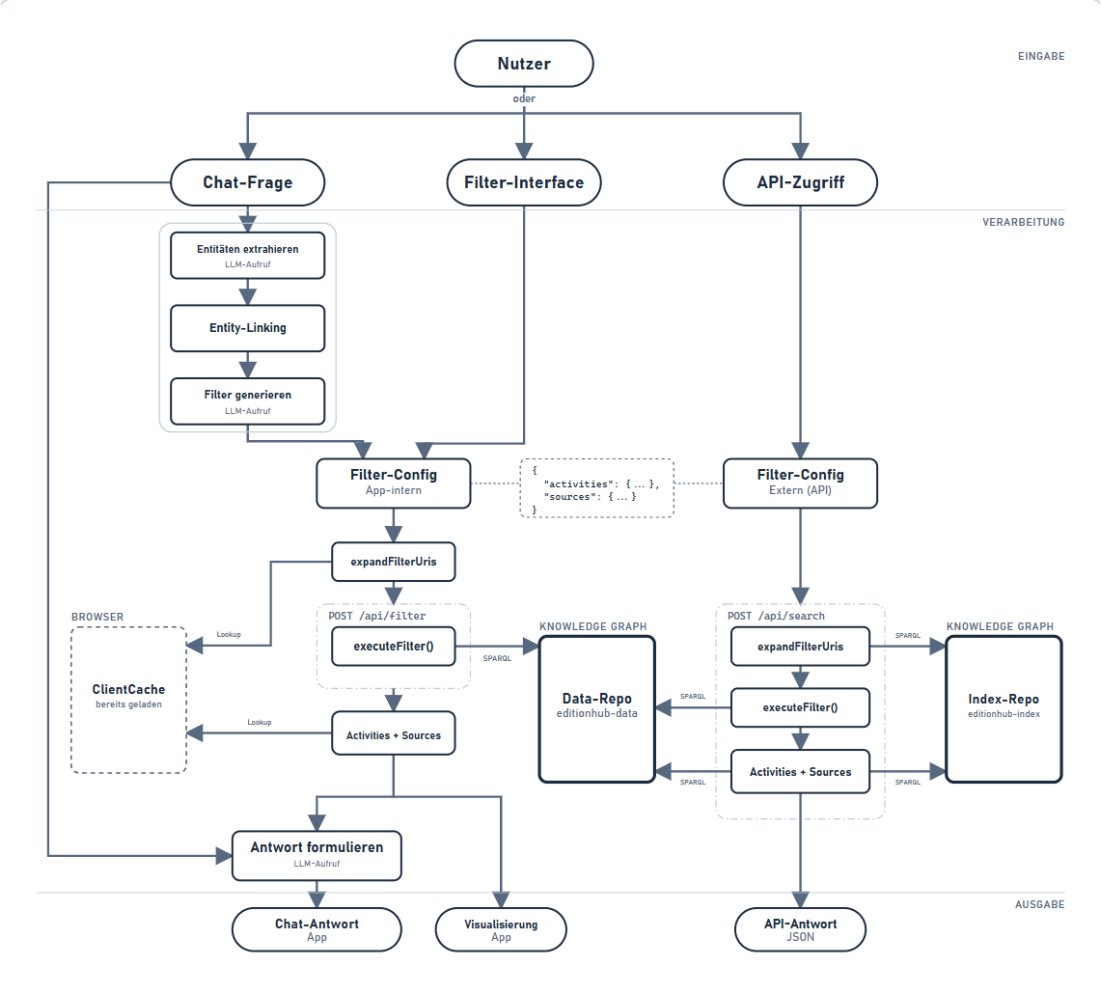
- Über eine API lässt sich der Datenbestand gezielt filtern und abfragen, um die Daten in anderen Kontexten und Projekten nachzunutzen.
- Über das grafische Filter-Interface auf der Plattform lässt sich der Datenbestand durchsuchen und auf verschiedene Arten visualisieren.
- Über einen KI-Chat auf der Plattform können Fragen gestellt werden, die ein Large Language Model (LLM) auf Grundlage der Plattformdaten beantwortet.

Aus technischer Perspektive basieren alle drei Zugangsmöglichkeiten auf einem gemeinsamen JSON-Format[6], in dem sich die gewünschte Filter-Konfiguration (Filter-Config) abbilden lässt und über das auf die Daten der Plattform zugegriffen wird (s. Grafik). Das Format beschreibt sowohl die Muster der gesuchten Aktivitäten — etwa Typ, Datum, Ort, Akteur, Objekte (materieller Kontext) und Bezüge (kognitive Inhalte) — als auch die zugehörigen Quellen (Typ, Edition, Archiv). Durch die Kombination mehrerer Aktivitäten — etwa Briefsendung und Briefempfang — miteinander und mit ihren Quellen lassen sich auch komplexe Zusammenhänge im Knowledge Graph gezielt abfragen.

Bei externen Datenabfragen über die API der Plattform https://historia.editionhub.org/api/search) werden zunächst die in der Filter-Config angegebenen Entitäten-URIs gegen die Index-Daten der

Plattform abgeglichen und um ihre sameAs-Aliasse erweitert. Entitäten (Personen, Orte etc.) können mit projektinternen URIs in der Plattform gespeichert werden. Der Index hält die Konkordanzen zwischen diesen projektinternen URIs und der jeweiligen kanonischen URI der Plattform vor. So lassen sich die Entitäten in projektübergreifenden Registern erfassen und weiterhin über ihre projektinternen URIs abfragen. Die erweiterte Filter-Config wird mittels executeFilter() in eine SPARQL-Anfrage an das Daten-Repo (Aktivitäten und Quellen) übersetzt, die die IDs der gesuchten Aktivitäten und Quellen zurückgibt. Diese IDs werden anschließend um die zugehörigen Informationen aus dem Daten- und Index-Repo ergänzt, sodass der Nutzer ein JSON mit den kompletten Daten der gesuchten Aktivitäten und Quellen erhält. (s. Grafik, rechte Seite)

Im Webinterface lassen sich über ein Filter-Panel Aktivitäten und ihre Quellen auswählen und miteinander kombinieren. Die Filter-Config wird zunächst durch einen clientseitigen Cache-Lookup angereichert, wobei die enthaltenen URIs um ihre sameAs-Aliasse erweitert werden. Nach der Ausführung von executeFilter() werden die zurückgegebenen IDs durch einen erneuten Cache-Lookup zu den vollständigen Daten aufgelöst. Beide Schritte laufen funktional analog zur Search-API ab, aus Performance-Gründen jedoch über einen clientseitigen Cache-Lookup statt über einen erneuten Server-Request — die benötigten Daten sind im Webinterface ohnehin bereits geladen. Alle Entitäten und Kategorien der abgefragten Aktivitäten und Quellen werden auf der Plattform in eigenen Registern übersichtlich gruppiert bereitgestellt. Die Aktivitäten werden sowohl auf einer Zeitleiste als auch auf einer Karte dargestellt. Darüber hinaus lassen sich die sozialen, semiotischen und semantischen Wissensnetzwerke mithilfe eigener Widgets visualisieren. (s. Grafik, Mitte) Alle Visualisie-



rungen sind miteinander verknüpft und ermöglichen eine interaktive Exploration der Daten aus unterschiedlichen Perspektiven. Durch eine gezielte Nutzung der Plattform kann bestimmten Forschungsfragen nachgegangen werden.

Abb.: Übersicht der drei Zugänge (API, Filter-Interface, KI-Chat) zum gemeinsamen Wissensgraphen.

Darüber hinaus lassen sich über ein Chat-Dialog natürlichsprachige Fragen an die Plattform stellen und mithilfe von LLM-based Knowledge Graph Question Answering (KGQA) beantworten lassen. Zunächst werden Entitäten extrahiert und per Entity-Linking aufgelöst, anschließend übersetzt ein LLM die Frage in eine Filter-Config. Von dort an läuft derselbe Prozess wie beim Filter-Interface: Die zurückgegebenen Daten gehen zusätzlich an das LLM, das auf ihrer Basis die Ursprungsfrage beantwortet. Filter und Visualisierung entstehen dabei parallel zur LLM-Antwort aus denselben Daten — so lassen sich KI-Antwort und Datenvisualisierung gegenseitig abgleichen und gemeinsam interpretieren. (s. Abb., linke Seite)

Das Projekt „PROPYLÄEN. Forschungsplattform zu Goethes Biographica“ wird im Rahmen des Akademienprogramms der Union der Deutschen Akademien der Wissenschaften gefördert und von der Klassik Stiftung Weimar, der Sächsischen Akademie der Wissenschaften zu Leipzig und der Akademie der Wissenschaften und der Literatur Mainz (2015–2039) durchgeführt. Die Arbeitsstellen befinden sich im Goethe- und Schiller-Archiv in Weimar sowie am Freien Deutschen Hochstift Frankfurt am Main.

Endnoten

[1] EditionHub:historia, siehe: https://historia.editionhub.org/
[2] CIDOC Conceptual Reference Model (CRM), siehe: https://cidoc-crm.org/sites/default/files/cidoc_crm_version_7.3.2.pdf
[3] PROPYLÄEN. Goethes Biographica, siehe: https://goethe-biographica.de/.
[4] Text Encoding Initiative (TEI), siehe: https://tei-c.org/
[5] Historical Activity and Source Exchange (HASE), siehe: https://editionhub.org/hase
[6] JavaScript Object Notation (JSON), siehe: https://www.json.org/json-en.html

Literatur

- Renn, Jürgen (Hrsg.). 2012. The Globalization of Knowledge in History. Berlin: Max-Planck-Gesellschaft zur Förderung der Wissenschaften.
- Renn, Jürgen, Dirk Wintergrün, Roberto Lolli, Manfred Laubichler und Matteo Valleriani. 2016. „Netzwerke als Wissensspeicher. Wissensrepräsentation im Internet“. In Die Zukunft der Wissensspeicher, herausgegeben von Jürgen Mittelstraß und Ulrich Rödiger, 35–79. München: UVK Verlagsgesellschaft.

AGATE meets KI – Einblicke in die experimentelle Öffnung des AGATE Knowledge Graphen mit LODEON

Linnaea Söhn, Akademie der Wissenschaften und der Literatur, Mainz; Martin Sievers, Akademie der Wissenschaften und der Literatur, Mainz; Torsten Schrade, Akademie der Wissenschaften und der Literatur, Mainz

Lightning Talk

Auf dem Jahresworkshop der AG eHumanities 2024 wurde AGATE (https://agate.academy) als FAIRes Informationssystem mit Datenschnittstelle vorgestellt – als Linked-Open-Data-Anwendung, deren Daten als RDF über eine Schnittstelle bereitgestellt und in einem Knowledge Graph über einen öffentlichen SPARQL-Endpunkt (https://agate.academy/sparql-endpoint.html) zugänglich gemacht werden (Söhn und Sievers 2024).

Es zeigte sich, dass auf dieser Grundlage Auswertungen zum Akademienprogramm vorgenommen werden können, die Aufschluss über Desiderate und Forschungspotenziale in der Grundlagenforschung geben – etwa durch strukturierte SPARQL-Abfragen zu geographischen oder zeitlichen Schwerpunkten musik- und editionswissenschaftlicher Projekte (so auch Söhn und Sievers 2025). Auch Frederike Neuber und Lou Klappenbach nutzten AGATE in ihrem Vortrag „Digitale Methoden für eine geschlechtergerechtere Forschung an den Akademien“ als Ressource, um ihre These eines Gender-Data-Gaps zu untermauern, und zeigten, wie Projektbeschreibungen und Titel von Forschungsprojekten reflexiven Aufschluss geben können (Neuber und Klappenbach 2024).

Beide Beispiele zeigen das Potenzial, das in den Daten steckt, machen jedoch auch deutlich, welche technischen und analytischen Voraussetzungen für dessen Hebung erfüllt sein müssen: Kenntnis der Datenstruktur, Abfragen mit SPARQL, eigene Auswertungen, gegebenenfalls Verarbeitungspipelines. Diese Voraussetzungen sind nicht bei allen Nutzer:innen gegeben – mit der Folge, dass ein erheblicher Teil des Potenzials von AGATE für Personen ohne SPARQL-Erfahrung ungenutzt bleibt. Zwar bietet AGATE auch ein grafisches Benutzerinterface mit umfangreichen Filtermöglichkeiten sowie eine Freitextsuche, doch lassen sich komplexere Fragen – etwa nach einem Gender Gap in Editionsprojekten – damit allein kaum beantworten. Jedes User Interface setzt zudem Designentscheidungen voraus: Welche Filter werden angeboten? Was ist im Suchindex enthalten? An welcher Stelle wird welches Projekt angezeigt? Somit stellt sich die Frage, wie sich die Auswertungsmöglichkeiten mit AGATE möglichst flexibel und mit gleichzeitig möglichst wenigen Restriktionen öffnen lassen – auch ohne SPARQL-Kenntnisse und ohne eigene Verarbeitungspipelines.

Dieser Herausforderung haben wir uns mit einem experimentellen, evidenzbasierten Ansatz gestellt. Dazu haben wir den AGATE Knowledge Graph an die im Rahmen von NFDI4Culture (https://nfdi4culture.de) entwickelte Forschungsanwendung LODEON (https://datastories.nfdi4culture.de/lodeon/) angeschlossen (siehe auch das Authoring Tutorial des Culture Data Story Laboratory (https://datastories.nfdi4culture.de/tutorial/)).

LODEON ist eine KI-gestützte Analyse- und Explorationsumgebung für Wissensgraphen, die an der Digitalen Akademie der Akademie der Wissenschaften und der Literatur | Mainz mit spezifischem Fokus auf Anwendungsfälle in den Geistes- und Kulturwissenschaften entwickelt wird. Mit einem besonderen Schwerpunkt auf Prinzipien digital souveräner Forschungsanwendungen (local-first, data minimisation, reproducibility, transparency) führt sie Graphdaten, narrative Elemente, evidenzbasierte Exploration und direkte KI-Unterstützung zusammen. Die KI-Assistenz übernimmt dabei eine zentrale Rolle: Über eine standardisierte Tool-Schnittstelle (MCP) greift sie kontrolliert auf Wis-

sensgraphen zu und unterstützt so nicht nur die natürlichsprachliche, explorative Erkundung, sondern erstellt auf Anfrage auch die für Auswertungen notwendigen strukturierten SPARQL-Abfragen sowie darauf basierende Visualisierungen.

Entscheidend ist dabei, dass diese KI-assistiert erzeugten Auswertungen keine Black-Box-Ergebnisse bleiben: Da die KI stets über klar definierte, profilgebundene Tools auf die Daten zugreift, ist jede Abfrage und jede Visualisierung unmittelbar an die zugrundeliegenden Daten in AGATE rückgebunden und lässt sich Schritt für Schritt nachvollziehen und überprüfen. Abgesichert wird dies durch konfigurierbare Competency Questions, die zugleich als kuratierte Beispiele für die KI-Assistenz und als ausführbare Benchmark-Fälle dienen. An ihnen lässt sich fortlaufend prüfen, ob und wie belastbar sich fachliche Fragestellungen für einen jeweiligen Wissensgraphen evidenzbasiert beantworten lassen – auch über Modell-, Prompt- oder Profiländerungen hinweg. Nutzer:innen werden so dabei unterstützt, mit Graphdaten zu interagieren, Erkenntnisse evidenzbasiert zu dokumentieren, visuelle und mediengestützte Interpretationen zu erstellen sowie ihre Arbeit nachhaltig zu teilen oder zu publizieren.

Zu den besonderen Funktionen von LODEON zählen insbesondere frei konfigurierbare Knowledge-Graph- und LLM-Modellprofile, ein visueller Query Builder auf Basis von Sparnatural (https://sparnatural.eu/), lokale Medien-Assets, ein IIIF-Viewer, ein dynamischer Graph-Explorer-Knoten sowie ein Verovio (https://www.verovio.org/)-Knoten, ergänzt um strukturierte Publikationsmetadaten mit Preview-, Export- und Publishing-Workflows.

Im Lightning Talk möchten wir unsere laufenden Experimente vorstellen und diskutieren: Wie müssen die Forschungsinformationen in AGATE strukturiert sein, um für eine KI gestützte Exploration besonders aufschlussreich zu sein? Welche Rolle spielen Graphdatenbank und Sprachmodellwahl? Und: Welche infrastrukturellen Voraussetzungen braucht es, damit KI Barrieren im Umgang mit strukturierten Daten senkt – oder errichtet sie stattdessen neue?

Bei der sprachlichen Ausarbeitung dieser Einreichung wurde unterstützend Claude Sonnet 5 genutzt. Die inhaltliche und konzeptionelle Verantwortung liegt vollständig bei den Autor:innen.

Literatur

- Neuber, F., und L. Klappenbach. „Digitale Methoden für eine geschlechtergerechtere Forschung an den Akademien“. Unveröffentlichter Vortrag, November 2024.
- Schrade, T. LODEON. V. 0.0.1. 2026. Veröffentlichung des Quellcodes in Vorbereitung.
- Söhn, L. C., und M. Sievers. „AGATE: Vom Nachweissystem für Projekte zum FAIRen Forschungsinformationssystem mit Datenschnittstelle“. Unveröffentlichter Vortrag, November 2024.
- Söhn, L. C., und M. Sievers. „AGATE: Kartierung Des Akademienprogramms Von Gestern Bis Heute“. Zenodo, August 29, 2025. https://doi.org/10.5281/zenodo.16997562.

Explorative Vernetzung verteilter Wissensquellen mit NAKAR

Samuel Schepp, Technische Hochschule Mittelhessen; Sebastian Enns, Technische Hochschule Mittelhessen; Andreas Kuczera, Technische Hochschule Mittelhessen

Lightning Talk

Geisteswissenschaftliche Forschungsdaten liegen häufig in unterschiedlichen, technisch und strukturell heterogenen Systemen vor. Digitale Editionen, Normdateien und offene Wissensbasen enthalten dabei jeweils eigene Informationen zu Personen, Orten, Werken, Ereignissen oder Quellen. Obwohl diese Entitäten häufig miteinander verbunden sind, befinden sich die relevanten Informationen

meist in voneinander getrennten Datenbeständen (Hyvönen 2023, 729ff.). Gemeinsame Identifikatoren, etwa aus der Gemeinsamen Normdatei (GND) oder aus Wikidata, ermöglichen zwar eine eindeutige Referenzierung, lösen jedoch nicht automatisch die technischen, syntaktischen und semantischen Unterschiede zwischen den Quellen (Jegan et al. 2023, 829ff.).

Ein bestehender Ansatz zur Zusammenführung verteilter Daten ist der Culture Knowledge Graph (CKG) der NFDI4Culture. Dabei werden ausgewählte Metadaten aus dezentralen Kulturerbe-Ressourcen übernommen, in RDF überführt und auf gemeinsame Ontologien abgebildet. Die vollständigen Forschungsdaten verbleiben bei den jeweiligen Anbietern (Tietz, Söhn, et al. 2025; Bruns et al. 2025). Dadurch können bereits integrierte Ressourcen gemeinsam durchsucht werden. Allerdings hängt der Informationsumfang des CKG davon ab, welche Datenquellen aufgenommen wurden und welche Metadaten diese bereitstellen. Informationen, die darüber hinausgehen, müssen bei Bedarf zusätzlich abgerufen und mit dem bestehenden Untersuchungskontext verbunden werden.

Eine Anwendung wie SHMARQL ermöglicht außerdem klickbare Navigation, SPARQL-Abfragen, Volltextsuche und narrative Data Stories mit dynamischen Visualisierungen (Tietz, Posthumus, et al. 2025). Sie unterstützt damit das Browsen vorhandener Knowledge Graphs. Allerdings fehlt ein Prozess, bei dem Forschende ausgehend vom aktuell sichtbaren Netzwerkausschnitt selbst fachliche Perspektiven auswählen, weitere Quellen gezielt abfragen und deren Ergebnisse unmittelbar in den laufenden Untersuchungskontext übernehmen können.

Damit ergeben sich drei zentrale Problemfelder. Erstens müssen Informationen aus unterschiedlichen Quellen bedarfsgerecht zusammengeführt werden (P1). Zweitens benötigen Forschende eine fachlich steuerbare Exploration, bei der sie während ihrer Untersuchung schrittweise neue Perspektiven und Datenquellen einbeziehen können (P2). Drittens müssen die entstehenden Zusammenhänge so dargestellt werden, dass sie trotz der Komplexität wissenschaftlich erkennbar und interpretierbar bleiben (P3). Gerade bei großen Wissensnetzwerken kann die Darstellung vieler Knoten und Beziehungen zum sogenannten Hairball-Effekt führen, bei dem relevante Pfade nur schwer erkennbar sind (Yoghourdjian et al. 2021). Auch bestehende visuelle Zugänge bieten häufig nur begrenzte Perspektiven und Analysewege (Battisti und Daquino 2026, 289ff.).

Konzept

An diesem Punkt setzt die Erweiterung von NAKAR an. NAKAR wurde als szenariobasierte Umgebung zur visuellen Exploration komplexer Wissensnetze entwickelt. Bereits zuvor unterstützte das System konfigurierbare Szenarien, Filter- und Drilldown-Funktionen sowie die schrittweise Erweiterung sichtbarer Teilgraphen (Schepp, Enns und Kuczera 2026). Die neue Erweiterung ermöglicht zusätzlich, verteilte Wissensquellen bedarfsgesteuert einzubinden. Die externen Daten werden dabei nicht dauerhaft in die Projektdatenbank integriert, sondern nur für den jeweiligen Untersuchungskontext abgefragt und temporär im Workspace dargestellt.

Das zentrale Element ist das Szenario. Es beschreibt eine fachliche Perspektive auf die Daten und legt unter anderem einen Ausgangspunkt, die einzubeziehenden Datenquellen und die auszuführenden Abfragen fest. Zusätzlich wird bestimmt, welche Knoten- und Beziehungstypen im Workspace angezeigt werden. Szenarien können gespeichert, verändert, erneut ausgeführt und miteinander kombiniert werden. Bei ihrer Ausführung sendet NAKAR passende Abfragen an mehrere Datenquellen. Lokale Forschungsdaten können beispielsweise über Cypher aus Neo4j abgefragt werden, während RDF-basierte Wissensquellen über SPARQL angesprochen werden. Die Originaldaten verbleiben in ihren jeweiligen Systemen; lediglich die für das Szenario relevanten Ergebnisse werden an NAKAR übertragen. Ein Adapter führt die Ergebnisse anschließend in einem gemeinsamen Labeled Property Graph (LPG) zusammen. Dieser stellt Informationen als typisierte Knoten und

Beziehungen mit zusätzlichen Eigenschaften dar. Dadurch können Ergebnisse aus verschiedenen Quellen innerhalb eines gemeinsamen Netzwerks untersucht werden. Gemeinsame Identifikatoren wie GND-Nummern oder Wikidata-IDs helfen dabei, zusammengehörige Entitäten zu erkennen und miteinander zu verbinden. Der entstehende Graph ist ausdrücklich nur eine temporäre Arbeitsansicht und keine neue zentrale Datenbank. Die ursprünglichen Quellen bleiben unverändert. Im NAKAR-Workspace kann der temporäre Graph jedoch direkt untersucht, gefiltert und durch weitere Szenarien erweitert werden. Besonders wichtig ist die Möglichkeit, Szenarien aufeinander aufzubauen. Erkennt NAKAR bei einem ausgewählten Knoten beispielsweise eine GND-Nummer oder Wikidata-ID, können passende Szenarien angeboten werden. Der vorhandene Identifikator wird automatisch als Parameter an die entsprechende Abfrage übergeben. Dadurch entsteht keine starre Abfolge von Untersuchungsschritten; vielmehr können Forschende abhängig von den jeweils gewonnenen Erkenntnissen weiterarbeiten. Unterstützt werden derzeit Neo4j-Datenbanken, generische SPARQL-Endpunkte und Wikidata. Weitere Quellentypen können über zusätzliche Adapter integriert werden.

Anwendung

Die praktische Funktionsweise zeigt sich unter anderem am Projekt „Korrespondenzen der Frühromantik“, das die Korrespondenzen der Jenaer und Berliner Frühromantik digital ediert und für literaturwissenschaftliche sowie netzwerkanalytische Untersuchungen erschließt (Breuer, Deicke und Burch 2022). Ein Knoten zu Friedrich Schlegel enthält beispielsweise eine Wikidata-ID. Dadurch kann NAKAR passende Wikidata-Szenarien anbieten, mit denen unter anderem Verwandtschaftsbeziehungen und Bildinformationen abgerufen werden. Die Ergebnisse werden anschließend in den temporären Graphen übernommen und gemeinsam mit den ursprünglichen Projektdaten dargestellt. Die Untersuchung kann somit von einem Projektknoten aus schrittweise in eine externe Wissensquelle erweitert werden.

Ein weiteres Beispiel stammt aus der digitalen Edition des Briefwechsels von Antje Brons (Kuczera und Voß 2023). Hier kann zunächst über eine Cypher-Abfrage ein bestimmter Text und die damit verbundene Personenentität Samuel Cramer aus der Publikationsumgebung der Edition abgerufen werden. Die Personenentität besitzt eine GND-Kennung, über die eine Verbindung zu einem GND-Knoten hergestellt wird. Diese Verbindung bildet wiederum den Übergang zu der entsprechenden Wikidata-Entität. Anschließend wird die GND-Kennung als Parameter für ein Wikidata-Szenario verwendet. Dieses ruft Samuel Cramer und dessen dort dokumentierte Familienbeziehungen ab. Die neu hinzugefügten Personen können wiederum als Ausgangspunkte für weitere Abfragen dienen. So ist beispielsweise eine weitere Untersuchung in Wikidata oder eine Rückkehr zu den Projektdaten möglich. Auf diese Weise entsteht ein schrittweiser Wechsel zwischen Editionsdaten und externen Wissensquellen, ohne dass die zugrunde liegenden Datenbestände verändert werden.

NAKAR wird außerdem in den Projekten „Religion und Naturwissenschaft im Umbruch: Sozinianische Briefwechsel und Inedita digital“ eingesetzt (Daugirdas und Kuczera 2026). Trotz unterschiedlicher Forschungsgegenstände und Datenmodelle verfügen all diese Projekte über Normdatenidentifikatoren, durch die ihre Entitäten mit GND, Wikidata und weiteren Wissensquellen verbunden werden können. Projektspezifische Szenarien ermöglichen dadurch eine gezielte Erweiterung des jeweiligen Untersuchungskontexts.

Neben vorkonfigurierten Szenarien können technisch versierte Nutzer auch eigene Abfragen an angebundene Quellen formulieren. Gleichzeitig begrenzt NAKAR die Darstellung auf den durch Szenarien, Abfragen und Nutzerentscheidungen bestimmten Ausschnitt. Herkunft und Typ der Elemente

bleiben durch Farben, Beschriftungen und eine Legende unterscheidbar. Dadurch verbindet das System vorstrukturierte Untersuchungsschritte mit offenen Explorationsmöglichkeiten.

Fazit

NAKAR adressiert die drei beschriebenen Problemfelder aufeinander abgestimmt. Erstens können verteilte Informationen bedarfsgerecht abgefragt und über gemeinsame Identifikatoren in einem temporären Graphen zusammengeführt werden (P1). Zweitens übersetzen Szenarien technische Abfragen in fachlich benannte Untersuchungsschritte und ermöglichen eine schrittweise Erweiterung des Untersuchungskontexts (P2). Drittens unterstützen Filter, Vorschauen und konfigurierbare Darstellungen dabei, den sichtbaren Ausschnitt zu begrenzen und relevante Beziehungen in komplexen Netzwerken besser erkennbar zu machen (P3).

Der Ansatz kompensiert jedoch weder die semantische Qualität der Modellierung von einzelnen Quellen noch ermöglicht es deren dauerhafte Integration. Seine Möglichkeiten hängen von verfügbaren Schnittstellen, geeigneten Szenarien und übereinstimmenden Identifikatoren ab. Die Auswahl und Interpretation der dargestellten Beziehungen bleibt Aufgabe der Forschenden.

Trotz dieser Einschränkungen zeigen die Anwendungsfälle, dass projektspezifische Forschungsdaten mit Informationen aus GND, Wikidata und weiteren Wissensquellen verbunden werden können, ohne diese zuvor vollständig in einer zentralen Datenbank zusammenzuführen. NAKAR unterstützt damit eine situative und schrittweise Vernetzung verteilter Wissensbestände. Der wesentliche Vorteil liegt darin, dass Forschende den Untersuchungskontext während des Arbeitsprozesses erweitern können, ohne die zugrunde liegenden Datenquellen zu verändern. Für die Zukunft sind zusätzliche Adapter sowie eine systematische Evaluation der Szenarien und Visualisierungen gemeinsam mit Fachwissenschaftlern vorgesehen.

Literatur

- Battisti, Tommaso, and Marilena Daquino. 2026. „Conservative by Design: Epistemic Limits in Digital Humanities Visualisation Practices.“ *Umanistica Digitale*, no. 24 (July): 289–321. https://doi.org/10.60923/issn.2532-8816/22921.
- Breuer, Ulrich, Aline Deicke, and Thomas Burch. 2022. *Korrespondenzen Der Frühromantik. Edition – Annotation – Netzwerkforschung*. https://gepris.dfg.de/gepris/projekt/470517871.
- Bruns, Sasha, Alexandra Büttner, Heike Fliegl, et al. 2025. „Data Integration into the Culture Knowledge Graph.“ Version 1.0.1. NFDI4Culture, May 16. https://doi.org/10.5281/ZENODO.15430381.
- Daugirdas, Kęstutis, and Andreas Kuczera. 2026. *Religion Und Naturwissenschaft Im Umbruch: Sozinianische Briefwechsel Und Inedita Digital*. https://sozinianer.de.
- Hyvönen, Eero. 2023. „Digital Humanities on the Semantic Web: Sampo Model and Portal Series.“ *Semantic Web* 14 (4): 729–44. https://doi.org/10.3233/SW-223034.
- Jegan, Robin, Leon Fruth, Tobias Gradl, and Andreas Henrich. 2023. *Integrating Access to Authority Data for Improved Interoperability of Research Data in the Digital Humanities*. https://doi.org/10.18420/BTW2023-54.
- Khoo, Christopher S. G., Eleanor A. L. Tan, Siam-Gek Ng, Chwee-Fong Chan, Michael Stanley-Baker, and Wei-Ning Cheng. 2024. „Knowledge Graph Visualization Interface for Digital Heritage Collections: Design Issues and Recommendations.“ *Information Technology and Libraries* 43 (1). https://doi.org/10.5860/ital.v43i1.16719.
- Kuczera, Andreas, and Klaas-Dieter Voß. 2023. *Das Ringen Um Integration Und Selbstbehauptung: Der Briefwechsel Der Ostfriesischen Mennonitin Antje Brons Geb. Cremer Ten Doornkaat (1810–1902)*. https://gepris.dfg.de/gepris/projekt/527629598.
- Nečaský, Martin, and Štěpán Stenchlák. 2022. „Interactive and Iterative Visual Exploration of Knowledge Graphs Based on Shareable and Reusable Visual Configurations.“ *Journal of Web Semantics* 73 (July): 100713. https://doi.org/10.1016/j.websem.2022.100713.

- Schepp, Samuel, Sebastian Enns, and Andreas Kuczera. 2026. *Digitales Entdecken Im Graph: Einfache Zugänge Zu Komplexen Netzwerken*. February 20. https://doi.org/10.5281/ZENODO.18696572.
- Tietz, Tabea, Etienne Posthumus, Linnaea Charlotte Söhn, et al. 2025. *Knowledge Representation and Discovery for Cultural Heritage Research Data with CTO and SHMARQL*. November 3. https://doi.org/10.5281/ZENODO.17509231.
- Tietz, Tabea, Linnaea Charlotte Söhn, Oleksandra Bruns, et al. 2025. *From Culture to Core: Integrating Cultural Heritage Data into Cross-Domain Research Infrastructures*. https://doi.org/10.5445/IR/1000186487.
- Yoghourdjian, Vahan, Yalong Yang, Tim Dwyer, Lee Lawrence, Michael Wybrow, and Kim Marriott. 2021. „Scalability of Network Visualisation from a Cognitive Load Perspective." *IEEE Transactions on Visualization and Computer Graphics* 27 (2): 1677–87. https://doi.org/10.1109/TVCG.2020.3030459.

Session 5: Forschungsdaten, Infrastrukturen & Zusammenarbeit

Vom Zettel zur Quelle: Abrufsystem für ein lexikographisches Zettelarchiv mit der Anwendung „Zettel-Ranker"

Piroska Lendvai, Digital Humanities Forschung und Entwicklung, Bayerische Akademie der Wissenschaften; Isabelle Mandrin, Mittellateinisches Wörterbuch, Bayerische Akademie der Wissenschaften

Vortrag

Das Mittellateinische Wörterbuch (MLW) stützt sich auf ein seit 1939 angelegtes Zettelarchiv von rund 1,5 Millionen Belegzetteln, deren Zitate für die Artikelarbeit in den maßgeblichen Editionen wiedergefunden werden müssen. Für gut aufbereitete Quellentexte leistet das die hauseigene, bislang nur intern genutzte Datenbank dMLW bereits; schwierig bleiben die Fälle, in denen der Zettel nach einer alten Ausgabe zitiert, die neue aber abweichende Seiten- und Zeilenzahlen und mitunter einen anderen Wortlaut bietet. Unsere Anwendung „Zettel-Ranker" schließt diese Lücke: Aus dem von OCR fehlerhaft erkannten Zitat eines Zettels findet sie die Belegstelle im Volltext und gibt sie mit Werk, Seite und Zeile zurück. Im Kern ist diese Erschließungsfunktion der „Zettel-Ranker" eine Suchmaschine über den Editionsvolltexten, technisch auf Elasticsearch aufgesetzt. Damit stellt sie die Maschinenlesbarkeit geisteswissenschaftlicher Daten dort her, wo sie bisher fehlte. Wir vergleichen klassische lexikalische Suche mit Methoden Künstlicher Intelligenz (hybride und rein semantische Suche über Satzvektoren). Auf einem aus autoritativen Datenbankverknüpfungen gewonnenen Goldstandard erreicht BM25 mit Wortüberlappung und Tippfehlertoleranz das beste Ergebnis (97,6 Prozent hit@1), die hybride Suche liegt gleichauf, ohne durch die Vektoren hinzuzugewinnen, und die rein semantische fällt deutlich ab. Der Grund liegt in der Aufgabenstruktur: Gesucht wird nicht ein ähnlicher Text, sondern derselbe Text in verrauschter Wiedergabe. Die eigentliche Grenze ist daher nicht die Suchqualität, sondern die Abdeckung des Korpus. Das System ist seit August 2026 im Probeeinsatz der MLW-Redaktion.

Einleitung

Das Mittellateinische Wörterbuch dokumentiert die lateinische Sprache des Mittelalters im deutschen Sprachraum für die Zeit von etwa 600 bis 1280. Als Belegwörterbuch zitiert es zu jeder Wortbedeutung die Originalbelege, die sie stützen. 1939 gegründet, erfolgte die Materialsammlung nicht digital, sondern durch ein Zettelarchiv: etwa 4.000 lateinische Texte, die meisten nicht vollständig ausgewertet, sondern gezielt nach aussagekräftigen Belegen exzerpiert.

Diese Vorauswahl ist der eigentliche Wert des Bestands und lässt sich nicht nachträglich erzeugen: Eine vollständige Textdatenbank zeigt für ein materialreiches Lemma tausende Stellen an, die niemand in nützlicher Frist durcharbeitet, während sich mit den etwa 200 lexikographisch vorausgewählten Zetteln zum selben Lemma zielgerichtet arbeiten lässt. Wer den Zettelkasten erschließt, erschließt also nicht Rohmaterial, sondern ein über Jahrzehnte gewachsenes fachliches Urteil.

Die Zettel sind nahezu vollständig als Bilddigitalisate mit der hausinternen Datenbank „dMLW“ verbunden, ebenso die gut maschinenlesbaren Quellentexte. Schwierig bleiben die ungünstigen Fälle: (1) Quellentexte, die sich durch das Erscheinen einer neuen, textkritischen Edition geändert haben: bei etwa einem Viertel der Zettel fehlt die klare Werkangabe, und eine Stellenangabe mit Seite und Zeile liegt nur bei 0,9 Prozent vor, sowie (2) ein aus OCR resultierender, abweichender Wortlaut der Zetteltexte: der Zitattext existiert nur als fehlerhafte Erkennung einer hand- oder maschinenschriftlichen Notiz.

Unser Beitrag verbindet zwei in der Praxis untrennbare Fragen: die Messung von Zugänglichkeit und die Gestaltung der Oberfläche, über die sie stattfindet.

Daten

Kennzahl	Wert
Zettel im Archiv, digitalisiert und mit dMLW verbunden	rund 1.500.000
exzerpierte Quelltexte insgesamt	rund 4.000
Zettel ohne klare Werkangabe	25 %
Zettel mit Stellenangabe (Seite, Zeile)	0,9 %
Werke als Volltext im Suchkorpus	904
Passagen im Produktionsindex	664.454
Werke mit autoritativer Verknüpfung	1.116
davon als Volltext vorhanden	55 %
von den 40 meistzitierten Werken durchsuchbar	18
Goldstandard: Belege / auswertbar	4.850 aus 596 Werken / 3.127

Für die fehlenden Texte bestehen sehr unterschiedliche Aussichten: 52 Werke liegen als frei lizenzierte TEI-Ausgabe vor, 505 als nicht offen lizenziertes Digitalisat bei dMGH, 95 anderweitig verzeichnet; für 1.876 ist kein digitaler Text bekannt.

Suchmethode

Der Zettel trägt Lemma, Kopfzeile mit Werkangabe, Zitat und Querverweise; diese Teile werden getrennt. Aus Lemma und Zettel wird ein Suchtext gewonnen, die Werkangabe auf eine Korpusdatei aufgelöst, die Passage im Volltext gesucht und mit Werk, Seite, Zeile und Kontext ausgegeben. Dazu gehört das Zusammenziehen von Wörtern, die am Zeilenende getrennt wurden; das geschieht für Suche und Anzeige gleichermaßen.

Im Kern des Zettel-Rankers arbeitet mit Elasticsearch eine Volltext-Suchmaschine. Indexiert wird nicht das Werk als Ganzes: Jede Quelldatei wird von einem gleitenden Fenster in überlappende Passagen zerlegt, das nach jeweils der halben Fensterlänge weiterrückt, sodass jedes Zitat vollständig in mindestens einer Passage steht. Die Fensterlänge entspricht der Größenordnung eines typischen Zettelzitats: Ein zu großes Fenster grenzt die Stelle ungenau ein, ein zu kleines zerschneidet das Zitat. Für den Textabgleich dient ein auf historisches Latein zugeschnittener Analyzer, der u/v und i/j zusammenführt sowie Groß-/Kleinschreibung und diakritische Zeichen normalisiert: orthographische Unterschiede zwischen Ausgaben werden im Index ausgeglichen.

Unser Goldstandard-Datensatz beruht nicht auf einer Handtranskription, die es für die Zettel nicht gibt: Die dMLW-Datenbank führt eine Spalte für korrigierte Erkennung, sie ist aber durchgehend leer. Autoritativ ist die hinterlegte Verknüpfung von Werk und Scanseite; ein Beleg wird aufgenommen, wenn sein Zitat zu mindestens 90 Prozent mit einem Textschnipsel dieses Werkes übereinstimmt. Die Goldstandard-Werte gelten also für ausrichtbare Zettel und sind insofern eine Obergrenze. Verglichen wurden gegen denselben Datensatz: Wortüberlappung mit Tippfehlertoleranz (BM25), eine hybride Variante aus Wörtern und Embeddings-Vektoren sowie rein vektorbasierte Suche. Die Satzvektoren haben wir mit Latin-BERT erzeugt (Bamman und Burns 2020), dem naheliegenden Kandidaten, weil es im Unterschied zu mehrsprachigen Modellen auf lateinischem Text trainiert wurde. Der Aufbau eines solchen Vektorindex über rund 600.000 Passagen dauert mehrere Stunden und benötigt eine GPU, gegenüber wenigen Minuten für den lexikalischen Index.

Ergebnisse

Gemessen wird mit hit@1 und hit@5 (richtige Stelle an erster bzw. unter den ersten fünf Positionen) sowie dem Mean Reciprocal Rank (MRR), der berücksichtigt, wie weit oben sie steht.

Verfahren	hit@1	hit@5	MRR
BM25, Wortüberlappung ab 50 %	97,6	99,3	0,984
hybrid (Wörter und Vektoren)	97,0	99,0	0,978
rein semantisch (Latin-BERT-Vektoren)	36,6	48,0	0,417

Die hybride Suche liegt mit 97,0 gegenüber 97,6 Prozent praktisch gleichauf mit BM25 (ein Unterschied von wenigen Belegen). Der Vektoranteil trägt also nichts bei, kostet aber die mehrfache Antwortzeit und den mehrstündigen Indexaufbau. Die rein semantische Suche fällt weit ab. Beides erklärt sich aus der Aufgabenstruktur: Für eine zeichenidentische Wiederauffindung genügt die lexikalische Überlappung, semantische Nähe stört eher. Das sagt nichts über die Qualität von Latin-BERT, sondern über die Passung von Modell und Aufgabe. In einem kontrollierten Gegentest schlossen wir aus, dass die Vektoren nur am Rauschen scheitern: Auch bei manuell korrigierten Anfragen übertrifft die hybride Suche BM25 nicht. Schlechte Texterkennung kostet dagegen 20 bis 70 Prozentpunkte (auf einer kleinen Stichprobe von 36 auswertbaren Paaren, daher nur als Illustration). Die dominante Fehlerquelle ist damit die Textqualität, nicht die Suchmethode.

Der Produktionsindex verzichtet folglich auf Vektoren, wird in wenigen Minuten gebaut und benötigt keine GPU. Das Auffinden von Werk und Seite ist damit für den Prototyp gelöst; offen bleibt die genaue Eingrenzung der Belegstelle innerhalb der gefundenen Passage: momentan überlappt der gefundene Textausschnitt die Goldstelle bei 41,0 Prozent der Treffer.

Die eigentliche Grenze ist die Quelltext-Abdeckung: Da nur etwa ein Viertel der zitierten Werke im Korpus liegt, ist „keine Quelle gefunden“ ein häufiges und legitimes Ergebnis, kein Systemfehler. Entscheidend ist dafür die Zitierhäufigkeit: Von den 40 meistzitierten Werken sind nur 18 durchsuchbar.

Unsere KI-Vorversuche für OCR der Zettel fielen ambivalent aus. Ein offen lizenziertes Bild-Text-Modell (NuMarkdown-8B-reasoning, auf Qwen2.5-VL-7B aufgebaut) verbessert auf einer gepaarten Stichprobe von 82 Zetteln zwei Maße: die mittlere Sauberkeit steigt von 76,7 auf 82,1 und die mittlere Treue der Kopfzeile von 61,7 auf 71,1. Sauberkeit misst ohne Referenztext, wie textähnlich das Erkannte überhaupt aussieht, als Anteil von Buchstaben und wortförmigen Zeichenketten abzüglich typischer Störzeichen; sie zeigt an, ob ein Text lesbar ist, nicht ob er richtig ist. Die Treue der Kopfzeile vergleicht dagegen gegen eine bekannte Referenz, das Werkkürzel aus der Datenbank, und misst als Zeichenähnlichkeit, wie genau die Kopfzeile tatsächlich getroffen wurde. Bei 65 der 82 Zettel ist die Erkennung sauberer.

Für die Suche bleibt das Ergebnis dennoch ausgeglichen: Auf den 1.259 neu erkannten Zetteln steigt der Recall, also der Anteil der Zettel, zu denen überhaupt eine Belegstelle gefunden wird, lediglich von 29,6 auf 31,9 Prozent: Die neue Erkennung findet zwar 122 Belegstellen, die die alte verfehlte, verliert aber 93, die sie gefunden hatte (wegen mehrspaltig aussehender Zettel). Bei 756 Zetteln findet keine der beiden Varianten eine Quelle, meist weil das zitierte Werk nicht im Korpus liegt; das begrenzt den Recall unabhängig von der Texterkennung. Sauberere Zeichen und bessere Auffindbarkeit korrelieren also nicht unbedingt in unseren Daten.

Die Zuordnung der Werkkürzel braucht ein fachliches Urteil der Redaktion. Von 1.544 Editionen mit Digitalisat-Link und 904 Korpuswerken waren anfangs nur 16 verknüpft; eine maschinell vorgeschlagene Konkordanz von 130 Zeilen, von der Redaktion vollständig durchgesehen, hob die Zahl auf 61. Der charakteristische Fehler solcher Vorschläge ist die Verwechslung zweier Werke desselben Autors, weil der Autorenteil den größten Teil der Zeichenähnlichkeit trägt.

Die Anwendung läuft seit Kurzem als Dienst im Browser mit Nutzenden aus der MLW-Redaktion.

Zusammenfassung und Ausblick

Der „Zettel-Ranker“ löst eine klar umrissene Aufgabe: Er findet zu einem verrauschten Zitat die Belegstelle in der Edition, mit 97,6 Prozent Trefferquote, und ergänzt dMLW dort, wo das Nachschlagen an abweichenden Ausgaben und fehlender Struktur scheitert. Der Weg zum Prototyp führte zu einem einfacheren System als erwartet, ohne Vektorindex und ohne GPU-Bedarf.

Als Nächstes stehen an: die Eingrenzung der Belegstelle innerhalb der gefundenen Passage; die Erweiterung des Korpus entlang der nach Artikelnennungen priorisierten Liste, wobei die 52 frei lizenzierten TEI-Ausgaben der aussichtsreichste Weg sind, weil TEI die Seitenzahlen mitführt; und der Anschluss an die Redaktionsdatenbank, damit bestätigte Belegstellen dorthin zurückgeschrieben werden können.

Literatur

- Bamman, David, und Patrick J. Burns. 2020. „Latin BERT: A Contextual Language Model for Classical Philology.“ arXiv:2009.10053. https://arxiv.org/abs/2009.10053. Verwendet in der Hugging-Face-Distribution https://huggingface.co/latincy/latin-bert
- „Elasticsearch“ (Version 9.4). https://www.elastic.co/elasticsearch
- „Mean reciprocal rank“. https://en.wikipedia.org/wiki/Mean_reciprocal_rank
- „Mittellateinisches Wörterbuch bis zum ausgehenden 13. Jahrhundert (MLW)“. 1959–. Herausgegeben von der Bayerischen Akademie der Wissenschaften. München: C. H. Beck. https://mlw.badw.de/das-projekt.html
- „dMGH“. https://www.mgh.de/de/mgh-digital/dmgh
- „BM25“. https://en.wikipedia.org/wiki/Okapi_BM25
- „NuMarkdown“. https://huggingface.co/numind/NuMarkdown-8B-reasoning

Lost in Translation? Über die Herausforderung der interdiszplinären Zusammenarbeit in Akademieprojekten.

Hannah Busch, CCeH / Universität zu Köln

Lightning Talk

In vielen Akademievorhaben werden historische Quellen (neu) erschlossen und ediert. Über die letzten Jahre hat sich auch hier ein 'digitaler Turn' vollzogen und neben der digitalen Bereitstellung der Daten gewinnt auch die quantitative Auswertung mehr und mehr an Bedeutung. (vgl. Borek et al. 2025) Diese Form der Langzeitprojekte ist immer auch einen Blick in die Konventionen und Traditionen der jeweiligen Fachdisziplinen. Dazu gehören die Editionspraxis, Auszeichnungssprachen und Standards in den digitalen Geisteswissenschaften sowie fachspezifischen Sprachkonventionen. Diese Ausgangslage, so die Hypothese, macht die Daten nicht unbedingt zugänglich für Personen außerhalb der eigenen Fachdisziplin. Doch welchen Einfluss hat das auf die Zusammenarbeit in den Akademieprojekten und die erfolgreiche Durchführung? Wie verändert es den Blick der involvierten Expert:innen auf die Strukturierung der Daten und welchen Einfluss hat das auf unsere projektinterne Kommunikation?

Der hier skizzierte Lightning Talk reflektiert die Herausforderungen der interdisziplinären Kollaborationen und die damit verbundenen Erwartungshaltung innerhalb der Projektteams. Als Beispiel dient das Projekt „Die Formierung Europas durch Überwindung der Spaltung im 12. Jahrhundert“ (BAdW/AWK 2023–2040), mit einem besonderen Schwerpunkt auf Datenbasis und die zentrale Forschungsfrage des Projekts. Das hier zu Grunde liegende Vorhaben erarbeitet eine neue und aktualisierte Übersicht der Quellenüberlieferung zum Alexandrinischen Schisma (1159–1177) in Regestenform. Es ist die erste systematische Sammlung seit dem 1888 erschienen und als Jaffé-Löwenfeld bekannten „Regesta Pontificum Romanorum“. Die Datenbasis im Projekt besteht zum einen aus gedruckten historischen Regestenwerken aus dem späten 18. und frühen 19. Jahrhundert, die für das Projekt digitalisiert, teilweise mit OCR neu erfasst und auf das Datenmodell des Projekts gemapped wurden.[1] Zum anderen aus Regesten- und Urkundenrepositorien, die ihre Daten in unterschiedlichen und nicht unbedingt standardisierten Formaten bereitstellen.[2] Neben den unterschiedlichen in den Projekten angewendeten Datenmodellen sind die Regesten auch mehrsprachig, die dominanten Sprachen sind bis heute Deutsch und Latein, im Projekt wird die Bereitstellung aller Regesten auf Englisch angestrebt, um die Internationalität zu gewährleisten. Dabei soll auch maschinelle Übersetzung Anwendung finden.

Anhand von zwei Beispielen aus der aktuellen Projektarbeit soll demonstriert werden, welche Herausforderungen sich speziell aus dem Alter der Quellen (analog und digital) sowie aus der Interdisziplinarität (besonders Diplomatik, mittelalterliche Geschichte und Digital Humanities) ergeben. Das erste Beispiel nimmt die Entwicklung des gemeinsamen Datenmodells in den Blick, bei dem ein besonderer Augenmerkt auf einer systematisierten Aufstellung der Typisierungskriterien der Quellenüberlieferungen liegt und wie diese sinnvoll mit den Rechtsakten, Ereignissen und Akteuren des Schismas in Bezug gesetzt werden können, um eine maschinelle Auswertung der Daten zu ermöglichen.

Als zweites Beispiel dient ein Proof of Concept, das in Rahmen einer Studienarbeit erarbeitet wurde. Dabei geht es um die Exploration der auf unser Datenmodell gemappten Rohdaten aus den oben beschriebenen Quellen. Das Ergebnis ist eine interaktive Visualisierung der päpstlichen Kommunikationsnetzwerke und ihrer räumlichen und zeitlichen Verteilung. Die D3-Visualisierung besteht aus drei Ansichten, einer geographischen Karte, einem Netzwerkgraphen der Austeller-Empfänger-Beziehungen und einem Zeitstrahl der Urkundenproduktion. Lassen sich dadurch Tendenzen sichtbar machen, an wen Papst Alexander III. und seine insgesamt drei Gegenpäpste Urkunden ausstellten? Können wir bereits auf Grundlage der noch nicht bereinigten und korrigierten Daten Obödienzwechsel visualisieren und lokalisieren? Kann die frühe Visualisierung uns darüber hinaus dabei unterstützen Unzulänglichkeiten in den Daten und unserem Datenmodell zu erkennen und zu optimieren?

Endnoten

[1] Zu nennen sind hier unter anderem die sogenannten Pontifizienbände sowie die Regestenbände der Reihe „Papsturkunden in“ (siehe Göttinger Papsturkundenwerk).
[2] Bereits importiert sind die Regesta Imperii online und die relevanten Urkundenregesten aus dem Bestand von papsturkunden.de.

Literatur

- Borek, Luise, Hannah Busch, Karoline Döring, Nora Ketschik, und Katharina Zeppezauer-Wachauer. 2025. „Einleitung. Schnittstelle Mediävistik. Kollaborationen der Mittelalterforschung im digitalen Zeitalter“. Das Mittelalter. Perspektiven mediävistischer Forschung 30 (1): 1–17. https://doi.org/10.17885/HEIUP.MIAL.2025.1.25117.
- Göttinger Papsturkundenwerk (Pius-Stiftung für Papsturkundenforschung). 2026. Verzeichnis erschienener und geplanter Regesten- und Editionsbände. https://adw-goe.de/forschung/forschungsprojekte-akademienprogramm/papsturkunden-des-fruehen-und-hohen-mittelalters-1/pius-stiftung-fuer-papsturkundenforschung/veroeffentlichungen/
- Jaffé, Philip, Hrsg. 1888. Regesta pontificum Romanorum ab condita ecclesia ad annum post Christum natum 1198, Bd. 2: ab a. MCXLIII ad a. MCXCVIII. Leipzig.

Wer darf Infrastruktur bauen? KI-gestützte Entwicklung einer föderierten Bibliographie-Synchronisation zwischen zwei Akademievorhaben

Luis Sáenz, Sächsische Akademie der Wissenschaften zu Leipzig (Leipzig Akkadian Dictionary)

Lightning Talk

Das Leipzig Akkadian Dictionary (LAD, Sächsische Akademie der Wissenschaften zu Leipzig) und die electronic Babylonian Library (eBL) sind zwei altorientalistische Vorhaben, die weitgehend den-

selben bibliographischen Bestand benötigen. Getragen wird die eBL-Plattform von der LMU München und der Bayerischen Akademie der Wissenschaften; auf ihr setzt auch deren Akademievorhaben Cuneiform Artefacts of Iraq in Context (CAIC) auf. Beide Vorhaben sind an den FAIR-Prinzipien ausgerichtet, unterscheiden sich technisch jedoch grundlegend: Das LAD ist ein digitales Wörterbuch in XML/TEI Lex-0 und bleibt so weit wie möglich Lex-0-konform; die eBL ist eine dynamische, als Webanwendung zugängliche Datenbank, die überwiegend mit JSON arbeitet. Die LAD-Bibliographie stammt aus dem Vorgängerprojekt Supplement to the Akkadian Dictionaries (SAD) und wurde zum Projektbeginn im Januar 2025 aus der gedruckten Vorlage in ein maschinenlesbares Format überführt. Nach Gesprächen mit dem eBL-Team fiel die Entscheidung, deren Bibliographie zu übernehmen, den eigenen Bestand damit abzugleichen und die dort fehlenden Titel beizusteuern. Um wiederholtes Konvertieren zwischen XML und CSL zu vermeiden, wird die Bibliographie seither durchgängig in CSL-JSON geführt, obwohl das LAD ansonsten ein XML-Projekt ist. Das LAD gibt an dieser Stelle also bewusst die formale Einheitlichkeit seiner Datenhaltung zugunsten der Interoperabilität auf: Ein von beiden Seiten geteiltes, etabliertes Format ist die Voraussetzung dafür, die Bestände überhaupt abgleichbar und nachnutzbar zu halten.

Schon wenige Wochen nach dem Import zeigte sich das Grundproblem: Der lokale Bestand war bereits wieder veraltet. Ein systematischer Abgleich machte zugleich das eigentliche Hindernis sichtbar: Beide Seiten hatten seit der Übernahme eigenständig weitergearbeitet, und ein Teil der Titel wurde beiderseits unter verschiedenen IDs geführt. Das gemeinsame Format garantierte also weder gemeinsame Identität noch Interoperabilität; die eigentliche Hürde liegt nicht im Datenformat, sondern in der Frage, wer IDs vergibt und nach welchen Regeln. Daraus entstand die Anforderung, die eBL regelmäßig und automatisch abzurufen.

Gewählt wurde ein föderiertes statt eines zentralisierten Modells. Die eBL ist Source of Truth und Distributionsknoten. Das LAD hält einen vollständigen lokalen Spiegel des Bestandes, sodass Lesezugriffe schnell und unabhängig von der Verfügbarkeit des Partnersystems bleiben. Jedes Projekt kann darüber hinaus eigene Einträge führen, die im gemeinsamen Bestand (noch) nicht enthalten sind: Der planmäßige Abgleich überschreibt lokale Kopien gemeinsamer Titel, lässt die Eigenbestände aber unangetastet. Schreibzugriffe laufen über einen serverseitigen Proxy mit Maschinen-Authentifizierung; die Duplikatsprüfung erfolgt serverseitig bei der eBL und blockiert Verdachtsfälle, statt sie stillschweigend anzulegen. De facto behält damit jedes Projekt seine eigene Bibliothek, kann den gemeinsamen Bestand aber nutzen und zu ihm beitragen. Für die Frage der Zugänglichkeit ist genau das der Kern: Ein Projekt muss seine Datenautonomie nicht aufgeben, um an einer gemeinsamen Ressource teilzuhaben.

Der Beitrag berichtet aus der Werkstatt, und dabei vor allem über die personellen Voraussetzungen, unter denen dieses System entstanden ist. Die eBL verfügt über professionelle Softwareentwicklung; die Partnerschnittstelle wurde dort von einem ausgebildeten Entwickler spezifiziert und implementiert. Auf LAD-Seite steht ein Altorientalist ohne formale Programmierausbildung, ein Quereinsteiger in digitale Methoden. Die gesamte LAD-seitige Entwicklung erfolgte KI-gestützt. Damit rückten Architekturfragen stärker in den Vordergrund als Fragen der konkreten Codierung: welches Datenhaltungsmodell zu wählen ist, über welche Schnittstellen, in welche Richtung und in welchem Takt die beiden Server Daten austauschen, welche Fehlerfälle abzufangen sind und wo die Grenze zwischen automatischer und menschlicher Entscheidung verläuft. Auch die Editions- und Verwaltungsoberfläche, über die die bibliographischen Daten in LAD kuratiert werden, ist auf diesem Weg entstanden.

Daraus ergibt sich die These des Beitrags. Zugänglichkeit wird überwiegend als Zugang zu Daten und Diensten diskutiert. Der geschilderte Fall legt eine zweite Dimension frei: den Zugang zur Fä-

higkeit, Infrastruktur überhaupt zu bauen und damit selbst Zugangsmöglichkeiten zu schaffen. Kleinere Vorhaben und Fächer ohne eigene Entwicklungsstellen sind bislang strukturell darauf verwiesen, entweder auf Anschluss zu warten oder ihre Autonomie abzugeben. KI-gestützte Entwicklung verschiebt diese Schwelle spürbar, einem Kooperationspartner mit professioneller Entwicklungsabteilung auf technisch gleicher Augenhöhe zu begegnen, Architekturoptionen zu bewerten und Schnittstellenverträge zu prüfen.

Entstanden ist damit zugleich ein Austauschmodell, das an den FAIR-Prinzipien ausgerichtet ist und nicht auf diese eine Partnerschaft beschränkt bleibt. Die Architektur steht weiteren Projekten desselben Fachs offen, und sie ließe sich ebenso in anderen Disziplinen nachbauen: Statt in jedem Vorhaben eine eigene Bibliographie zu pflegen, könnte schrittweise ein belastbarer gemeinsamer bibliographischer Fachbestand entstehen, an dem jedes digitale Projekt teilhaben kann, ohne seine Eigenständigkeit aufzugeben. Die Implementierung wird derzeit von beiden Seiten intensiv erprobt; bis zur Konferenz steht daher nicht ein Vorhaben zur Vorstellung, sondern ein im Regelbetrieb bewährtes System.

Literatur

- Citation Style Language. n.d. „Citation Style Language Specification.“ Zugriff am 4. August 2026. https://citationstyles.org.
- Cuneiform Artefacts of Iraq in Context (CAIC). n.d. Bayerische Akademie der Wissenschaften. Zugriff am 4. August 2026. https://caic.badw.de.
- Electronic Babylonian Library (eBL). n.d. Ludwig-Maximilians-Universität München und Bayerische Akademie der Wissenschaften. Zugriff am 4. August 2026. https://www.ebl.lmu.de.
- Leipzig Akkadian Dictionary. n.d. Sächsische Akademie der Wissenschaften zu Leipzig. Zugriff am 4. August 2026. https://lad.saw-leipzig.de.
- Streck, Michael P., Hrsg. 2018–. Supplement to the Akkadian Dictionaries. Leipziger Altorientalistische Studien 7. Wiesbaden: Harrassowitz.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, u. a. 2016. „The FAIR Guiding Principles for Scientific Data Management and Stewardship.“ Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18.

Anonym Gebärden? Privatsphäre im Hamburger Korpus der Deutschen Gebärdensprache

Marc Schulder, Institut für Deutsche Gebärdensprache und Kommunikation Gehörloser, Universität Hamburg; Thomas Hanke, Institut für Deutsche Gebärdensprache und Kommunikation Gehörloser, Universität Hamburg

Lightning Talk

Das DGS-Korpus ist ein Referenzkorpus der Deutschen Gebärdensprache (DGS) (Hanke et al, 2020). Bestehend aus Videoaufnahmen von Gesprächspaaren ist es sowohl Grundlage für sprachwissenschaftliche Forschung als auch Zeitzeugendokument über das Leben tauber Personen in Deutschland. Teilnehmende berichten in den Aufnahmen frei über ihre persönlichen Lebenserfahrungen. Sie geben daher in den Aufnahmen mitunter sehr persönliche Informationen preis. Auch gehören die Teilnehmenden einer sprachlichen und kulturellen Minderheit an, die historisch systematisch unterdrückt wurde und auch heute noch vielen Barrieren ausgesetzt sind. Es liegt daher in der Verantwortung des Korpusprojekts, die Privatsphäre der Teilnehmenden so gut wie möglich zu schützen (Schulder und Hanke, 2022).

Das DGS-Korpusprojekt (2009–2027) ist ein an der Universität Hamburg durchgeführtes Langzeitvorhaben der Akademie der Wissenschaften in Hamburg. Teil seiner neunzehnjährigen Laufzeit ist die Durchführung zweier Datenerhebungen, die erste im Zeitraum 2010–2012, die zweite derzeit laufend. Auf Basis der sprachwissenschaftlichen Annotation dieser Aufnahmen entsteht das DGS-Korpus. Von den 560 Stunden Gesprächsmaterial der ersten Erhebung wurden über 50 Stunden bereits als öffentlich einsehbares Teilkorpus publiziert. Die Einschränkung der veröffentlichten Datenmenge bedingt sich durch die hierfür vorgesehenen Qualitätsstandards, die unter anderem bedeutungsverzerrende Fehlannotationen und die Preisgabe personenbezogener Daten verhindern sollen. Die öffentlichen Daten stehen außerdem unter strengen Lizenzbedingungen, die unter anderem kommerzielle Nutzung verbieten und die Nutzung der Annotationsdaten auf sprachwissenschaftliche Forschung begrenzen, um sicherzustellen, dass ihre Nutzung nur Zwecken zugute kommt, die im Sinne der Teilnehmenden und der Gehörlosengemeinschaften sind.

Auf der Stufe der Metadaten entspricht der Schutz personenbezogener Daten der allgemeinen guten Praxis für Daten mit menschlichen Teilnehmenden. Informationen werden nur veröffentlicht, wenn sie für die Nutzung des Datensatzes vonnöten sind, und werden in ihrer Granularität angepasst, um keine eindeutige Identifikation zu ermöglichen. Hierbei spielt die Bevölkerungsgröße eine wichtige Rolle. Es leben geschätzt 80,000 taube Menschen in Deutschlande, Informationen müssen also teils deutlich mehr vergröbert werden als bei Datensätzen der Gesamtbevölkerung. Ziel des DGS-Korpus ist unter anderem, die Erforschung von Sprachwandel, geschlechterspezifischer Sprache und regionalen Dialekten zu ermöglichen. Entsprechend sind Geschlecht, Alter und Herkunftsregion der Teilnehmenden Teil der öffentlichen Metadaten, allerdings verallgemeinert auf Generationskohorten (je fünfzehn Jahre abdeckend) und einer Unterteilung Deutschlands in dreizehn Regionen.

Bei den eigentlichen Daten des Korpus gestaltet sich der Schutz der Identität der Teilnehmenden allerdings deutlich schwieriger als bei vielen lautsprachlichen Korpora (Isard, 2020). DGS ist eine visuell-gestische Sprache, ihre vollständige Repräsentation erfordert bewegte Körper, bei medialer Darstellung also Video. Eine alltagstaugliche Schriftform für DGS gibt es nicht. Eine Übertragung ins Deutsche entspricht der Übersetzung in eine Fremdsprache, ist also als Primärdatenform ungeeignet. In den meisten Fällen bedeutet dies, dass bei einer gebärdeten Aussage auch immer die Person sichtbar sein muss, die die Aussage trifft. Gebärdensprachkorpora wie das DGS-Korpus müssen daher in Kauf nehmen, dass ihre Veröffentlichung auch immer die Preisgabe der visuellen Identität ihrer Teilnehmenden mit sich zieht. In den Datenerhebungen des DGS-Korpus werden Teilnehmenden daher als Teil der informierten Einwilligung über diesen Umstand aufgeklärt und ihnen die Möglichkeit eingeräumt, ihre Daten auf den nicht-öffentliche Teil des Korpus zu beschränken.

Auch wenn Teilnehmende in den Aufnahmen visuell identifizierbar sind, so gilt es trotzdem, die sonstigen personenbezogenen Daten von ihnen und Dritten zu schützen. Erwähnungen von Namen, Altersdaten und Wohnorten sowie Beschreibungen, aus denen diese Informationen hergeleitet werden können, werden anonymisiert. Hierfür müssen bedeutungstragende Teile des Videomaterials geschwärzt und Nennungen in dazugehörigen Annotationen und Übersetzungen durch Platzhalter ersetzt werden.

Die beschriebene Vorgehensweise des DGS-Korpus stellt die bisher als gute Praxis angesehene Vorgehensweise bei der Erstellung von Gebärdensprachkorpora dar. Angesichts der rasanten Entwicklungen der letzten Jahre bei Sprachtechnologien und Bildverarbeitung ist jedoch geboten, auch diese Praxis regelmäßig zu re-evaluieren, sowohl auf neue Potenziale als auf neue Hürden. So wächst das Interesse an Gebärdensprachdaten als Trainingsmaterial für generative Modelle stetig, nicht nur für Sprachtechnologien. Vor allem bei kommerziellen Akteuren ist zudem zu beobachten, dass das bewusste Ignorieren von Nutzungsbedingungen immer mehr zur gängigen Praxis wird.

Korpusprojekte müssen sich also fragen, ob die öffentliche Bereitstellung ihre Daten weiterhin verantwortbar ist, und vor allem bei neuen Erhebungen damit rechnen, dass auch potenzielle Teilnehmende der Veröffentlichung ihrer Daten zunehmend kritisch gegenüberstehen könnten.

Ein möglicher Schritt in der näheren Zukunft könnte sein, Teilnehmende mithilfe von generativen Modellen durch photorealistische Avatare zu ersetzen. Derzeitige Modelle haben noch deutliche Schwächen vor allem bei schnellen Handbewegungen und exaktem Mundbild, beides wichtige Faktoren für Gebärdensprachen. Sollte es in Zukunft jedoch möglich werden, solche sinnentstellenden Generierungsartefakte auszumerzen, wäre denkbar, öffentliche Avatar-Vorschaufassung anzubieten, während die Originaldaten hinter Zugriffsschranken gesichert werden. Selbst dann stehen weiterhin viele Fragen im Raum. Welche Forschungsfragen sind mit Avatar-Fassungen noch beantwortbar? Wie stehen Teilnehmende dazu, visuell ersetzt zu werden? Können solche Avatare mehr Akzeptanz in Gehörlosengemeinschaften finden als bisherige Modelle? Und wie anonym sind diese Fassungen wirklich? Denn selbst auf Strichfiguren heruntergebrochen erkennen fließend gebärdende Personen einander oft an ihrem idiosynkratischen Sprachstil (Battisti et al. 2024).

Literatur

- Hanke, Thomas, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. ‘Extending the Public DGS Corpus in Size and Depth’. Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives (Marseille, France), May 16, 75–82. https://www.sign-lang.uni-hamburg.de/lrec/pub/20016.html.
- Schulder, Marc, und Thomas Hanke. 2022. „How to Be FAIR When You CARE: The DGS Corpus as a Case Study of Open Science Resources for Minority Languages“. In Proceedings of the 13th International Conference on Language Resources and Evaluation, herausgegeben von Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, u. a. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.18.
- Isard, Amy. 2020. ‘Approaches to the Anonymisation of Sign Language Corpora’. In Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives, edited by Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch. European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/pub/20037.pdf.
- Battisti, Alessia, Emma van den Bold, Anne Göhring, Franz Holzknecht, and Sarah Ebling. 2024. ‘Person Identification from Pose Estimates in Sign Language’. In Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources, edited by Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Johanna Mesch, and Marc Schulder. {ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL)}. https://www.sign-lang.uni-hamburg.de/lrec/pub/24025.pdf.

Robustheitstests historischer Briefnetzwerke dank offener Forschungsdaten aus correspSearch, dem PROPYLÄEN-Projekt und dem Goethe- und Schiller-Archiv

Dr. Martin Prell, Sächsische Akademie der Wissenschaften zu Leipzig, Arbeitsstelle im Goethe- und Schiller-Archiv, Klassik Stiftung Weimar

Lightning Talk

Zielstellung

Nach dem Sozialhistoriker Charles Wetherell ist „incomplete historical record“ einer der Hauptgründe für die zurückhaltende Anwendung quantitativer, computationeller Methoden durch Historiker:innen (Wetherell 1998, 125). Dahinter steht noch immer häufig die Befürchtung, dass Überlieferungslücken netzwerkanalytische Ergebnisse so stark verzerren, dass daraus keine belastbaren historischen Erkenntnisse gewonnen werden können. Der Vortrag prüft diese Annahme empirisch-quantitativ anhand umfangreicher, offen zugänglicher Briefdaten. Netzwerkdatensätze aus correspSearch, dem PROPYLÄEN-Akademienprojekt und dem Goethe- und Schiller-Archiv werden dafür mit einer von Yann C. Ryan und Sebastian E. Ahnert entwickelten R/Shiny-Anwendung computationellen Robustheitstests unterzogen. Dabei werden die Auswirkungen unterschiedlicher Arten und Umfänge von Unvollständigkeit auf wichtige Netzwerkmetriken simuliert. Dies erlaubt netzwerkanalytische Rückschlüsse unterschiedlicher Art: Im Fall von correspSearch ermöglicht dies Aussagen darüber, wie zuverlässig die dort verfügbaren Daten vor dem Hintergrund einer stetig wachsenden Datenbank bereits jetzt sind. Im Fall der Briefdaten des PROPYLÄEN-Projektes und des Goethe- und Schiller-Archivs wiederum kann daraus abgeleitet werden, wie robust diese kuratierten Daten aus historisch gewachsenen Archiven gegenüber Missing Data angesichts zwangsläufiger Überlieferungslücken sind. Da alle Tests auf offen publizierten Briefforschungsdaten stattfinden, demonstriert die Analyse, welche Art und Qualität sowie welchen Umfang offen publizierte Daten aufweisen sollten, um netzwerkanalytischen Anforderungen zu genügen.

Forschungsstand und Methode

Historische Netzwerke sind aufgrund von Modellierungsentscheidungen, fragmentarischer Überlieferung und nicht schriftlich dokumentierter Kommunikation grundsätzlich unvollständig. In Korrespondenznetzwerken können Briefe, Korrespondenzpartner:innen oder ganze Briefkonvolute und Bestände fehlen. Da Netzwerkmetriken wie Zentralitäten relational berechnet werden, kann bereits der Verlust einzelner Elemente die Gesamtnetzwerkstruktur verändern. Ob und in welchem Maße dies geschieht, lässt sich jedoch nicht pauschal bestimmen, sondern muss für das jeweilige Netzwerk ermittelt werden.

Trotz der grundlegenden Bedeutung von Missing Data für historische Netzwerkanalysen wurde ihr Einfluss bislang selten systematisch untersucht (Ryan und Ahnert 2021; Valeriola 2021; Düring 2016). Die formale historische Netzwerkforschung begegnet fehlenden Daten bislang vor allem durch Strategien, die Verzerrungen zwar sichtbar machen, jedoch meist keine quantitative Aussage darüber erlauben, wie stark fehlende Daten konkrete Netzwerkstrukturen beeinflussen. Hierbei können Robustheitstests helfen. Diese simulieren den Verlust von Kanten oder Knoten und vergleichen die dabei entstehende Netzwerkstruktur mit der des Ausgangsnetzwerks. Die von den Netzwerkforschern Ryan und Ahnert entwickelte R/Shiny-Anwendung entfernt dazu sukzessive zufällige Kanten und misst die Übereinstimmung der sich verändernden Zentralitätsrankings mittels Spearman-Rangkorrelation (ρ)[1]. Je höher ρ, desto stärker stimmt das Zentralitätsranking der verbliebenen Knoten im reduzierten Netzwerk mit demjenigen im ‚vollständigen‘ Netzwerk überein. Auf diese Weise lässt

sich bestimmen, bis zu welchem Umfang und bei welchen Formen des Datenverlusts zentrale strukturelle Aussagen stabil bleiben.

Anwendung auf drei offene Netzwerkdatensätze

Das Tool wurde von Ryan und Ahnert an unvollständigen historischen Korrespondenznetzwerken erprobt und beide Autoren weisen darauf hin, dass sich mit ihm grundsätzlich beliebige Netzwerke prüfen lassen. Zugleich stellen die Anwendung ihres Tools und die Übertragbarkeit ihrer Robustheitsbefunde auf andersartige Netzwerke aber eine ausdrücklich offene methodische Frage dar (Ryan und Ahnert 2021, 68 und 83). Daher werden im Vortrag die Ergebnisse der Anwendung des Tools auf drei historische Korrespondenznetzwerke unterschiedlicher Genese, Größe und Struktur vorgestellt:

Diese umfassen die Netzwerke, die sich aus den Briefbeständen des Goethe- und Schiller-Archivs, den Briefen an Goethe des gleichnamigen Teilprojektes des PROPYLÄEN-Akademienvorhabens sowie den in correspSearch aggregierten Briefmetadaten (re)konstruieren lassen. Anhand der rund 375.000 Briefe und Briefbestände des Goethe- und Schiller-Archivs[2] lässt sich insbesondere untersuchen, wie sich unterschiedliche Disambiguierungs- und Normierungsstände eines historisch gewachsenen Archivs als eine besondere Art von Missing Data auf Zentralitätsrankings auswirken. Aus den 15.110 editorisch erschlossenen, handschriftlichen Briefe an Johann Wolfgang von Goethe der Jahre 1762 bis 1824[3] wird wiederum ein Kookkurrenznetzwerk erwähnter Personen erzeugt und die Robustheit dieses Netzwerks vor dem Hintergrund mehrfacher Autodafés Goethes analysiert. Die standardisierten und umfangreichen correspSearch-Daten mit derzeit 425.367 verfügbaren Briefversionen und den darin enthaltenen 55.869 Personen[4] ermöglichen nicht nur eine Prüfung der Skalierbarkeit des Verfahrens, sondern auch Vorhersagen darüber, wie sich die Zentralitätsrankings verhalten, wenn in Zukunft weitere Metadaten hinzukommen.

Für die drei Netzwerke werden zudem verschiedene Formen unvollständiger Überlieferung simuliert. Dazu zählen der Verlust einzelner Kanten, das Fehlen vollständiger Briefe oder Jahrgänge, der Ausfall einzelner Akteur:innen sowie die Entfernung ganzer Faszikel oder Bestände. Auf diese Weise lässt sich untersuchen, wie Netzwerkmodell, zufällige und gezielte Verlustmechanismen und verwendete Metrik gemeinsam die Stabilität netzwerkanalytischer Befunde beeinflussen. Daraus lassen sich wiederum Empfehlungen bezüglich Umfang, Art und Qualität bei der Erstellung und Publikation offener Daten für deren Nachnutzung im Rahmen von Netzwerkanalysen formulieren.

Endnoten

[1] Die Online-Testversion des Tools mit Kanten- und Iterationsbeschränkung ist unter https://networkingarchives.shinyapps.io/network_robustness_tool erreichbar. Der Progammcode für die installierbare R/Shiny-Anwendung ist hier zu finden: https://github.com/networkingarchives/network_robustness_ca_article.
[2] Vgl. die Ergebnisliste der Briefsuche innerhalb der Archivdatenbank des Goethe- und Schiller-Archivs unter https://ores.klassik-stiftung.de/ords/f?p=401:230:2634226426828.
[3] Vgl. die frei publizierten Metadaten aller Briefe an Goethe der Jahre 1762–1824 im CMIF-Format unter https://doi.org/10.5281/zenodo.14998880.
[4] Vgl. die Statistik auf https://correspsearch.net/ (Stand: 26. August 2026).

Literatur

- Düring, Marten. 2016. How reliable are centrality measures for data collected from fragmentary and heterogeneous historical sources? A case study. In: The connected past. Challenges to network studies in archaeology and history., hg. von Tom Brughmans, Anna Collar, und Fiona S. Coward, 85–101.

- Ryan, Yann C. und Sebastian E. Ahnert. 2021. The measure of the archive. The robustness of network analysis in early modern correspondence. Journal of cultural analytics 6 (3): 57–88. doi:10.22148/001c.25943.
- Valeriola, Sébastien de. 2021. Can historians trust centrality? Historical network analysis and centrality metrics robustness. Journal of historical network research 6 (1): 85–125. doi:10.25517/jhnr.v6i1.105.
- Wetherell, Charles. 1998. Historical social network analysis. International review of social history 43 (6): 125–144.

Crowdsourcing-Plattformen als Projekt-Inkubatoren vor der Forschungsförderung

Julian Jarosch, Akademie der Wissenschaften und der Literatur Mainz

Lightning Talk

Anhand eines lexikografischen Referenzprojekts ohne Forschungsförderung wird die Nutzung der kostenfreien und offenen Plattformen Wikisource und OpenStreetMap als Arbeitsumgebung und Datenspeicher umrissen und evaluiert.

Projektumriss

Die hier vorgestellten Aspekte des Referenz-Vorhabens sind die Retrodigitalisierung von Volltexten im Wikipedia-Schwesterprojekt Wikisource (https://fr.wikisource.org), und der Verknüpfung von Geo-Daten in der Online-Geodatenbank OpenStreetMap (https://www.openstreetmap.org).

Datengrundlage des Referenzprojekts

Das vorgestellte Referenzprojekt fokussiert sich auf lexikografische Daten, die der französische Erkundungsreisende und Mönch Charles de Foucauld (1858–1916) ca. in den Jahren 1905–1916 (Chatelard, 1995) in der zentralen Sahara gesammelt und erstellt hat. Zentral in seinem linguistischen Werk sind seine Wörterbücher der Tuareg-Varietät Tamahaq, gesprochen im Ahaggar und Nachbarregionen (heute in Algerien und Libyen). Foucaulds Lexikografie wirkt bis heute in der Forschungsgeschichte zum Tamahaq und schlägt sich als fundamental in aktuellen Publikationen (Ritter, 2009) nieder. Gleichzeitig ist ein Hemmschuh der Rezeption, dass sein zentrales, allgemeinsprachliches Wörterbuch mit 2028 Manuskriptseiten (Foucauld, 1951–1952) nie gesetzt wurde, sondern als Faksimile des druckfertigen Manuskripts erschienen ist. Die Nutzung des Wörterbuchs setzt also notwendigerweise die Konsultation der Handschrift des Autors voraus.

Neben dem allgemeinen Wörterbuch ist auch Foucaulds Eigennamen-Wörterbuch (Foucauld, 1940) in das Retrodigitalisierungsvorhaben einbezogen. Einen substanziellen Anteil des Eigennamen-Wörterbuchs machen die Ortsnamen aus, die neben linguistischen auch geografische Informationen enthalten. In diesem Sinne handelt es sich strukturell um ein gazetteer – eine Liste individueller Orte – und nicht nur um ein Ortsnamenbuch, das nach linguistisch einzigartigen Namen aufgebaut wäre. Durch diesen Aspekt verknüpft sich die Retrodigitalisierungsarbeit von den Text- und Sprachdaten auch zu Geo-Daten.

Ein sehr häufig im Eigennamen-Wörterbuch referenzierter Typ von Geografikum sind Trockenflusstäler. Deren linienförmige Geometrie ist besonders ungeeignet für eine pragmatisch vereinfachte Datenhaltung als einzelne Punktkoordinate. Forscherisch ist es also deutlich interessanter, sie in einem linearen Datentyp zu erfassen – was de facto den Einsatz einer Geoinformations-Datenbank nach sich zieht.

Die Retrodigitalisierung des Textes des umfangreichen allgemeinsprachlichen Wörterbuchs (2028 Manuskriptseiten) wurde 2010 begonnen, steht bei ca. 25 % Fortschritt bezogen auf Rohtranskriptionen, und pausiert derzeit überwiegend. Das Eigennamen-Wörterbuch wurde 2016–2018 retrodigitalisiert.

Wikisource

Wikisource bietet für die Retrodigitalisierung von Volltexten einen umgrenzten Spielraum in der Ausgestaltung der Erfassungs- und Editionsrichtlinien, die für ein Buch angewendet werden. Über diese Ausgestaltung hat konventionellerweise die Person Verfügungsfreiheit, die die Retrodigitalisierung anstößt und / oder durchführt. Der Gestaltungsspielraum ist auf technischer und auf normativer Ebene in einer fortlaufenden Abwägung mit dem Rahmen der gesamten Plattform. Die technische Ausstattung der Plattform ist durch die Erweiterung des Grundsystems (MediaWiki) um spezifische Extensions auf die Unterstützung des Datenmanagements und des Digitalisierungsworkflows ausgerichtet:

Die Datenhaltung baut auf die Einzelseite auf, was sich praktisch als ausreichend atomare Grundeinheit der Datenpflege bewährt. Die Präsentation der Texte in semantisch bestimmten Einheiten – wie Kapiteln oder sonstigen Abschnitten – wird von der Plattform ermöglicht und konfigurierbar gemacht.

Der Workflow der Digitalisierungsschritte wird von der Plattform auf Einzelseitenebene unterstützt durch Protokollierung und Visualisierung vordefinierter Arbeitsschritte wie Ersttranskription, Korrektur und Zweitkorrektur nach 4-Augen-Prinzip.

OpenStreetMap

OpenStreetMap kann als praktisch unbegrenzt in der Menge und Granularität der erfassbaren Geodaten (Punkte, Linien, Flächen etc.) charakterisiert werden. Allerdings unterliegt die normative Auswahl der typisierenden und beschreibenden Informationen an jeder Geometrie Eingrenzungen. Konkreter und beispielhaft ausgedrückt: OpenStreetMap begrenzt weder, wie viele Trockenflusstäler in die Datenbank eingetragen werden dürfen, noch mit welcher detaillierten Genauigkeit (Anzahl der Stützpunkte) jedes Tal modelliert werden darf. Andererseits wird durch Konsens in der Community reguliert, mit welchen kategorialen Angaben Trockenflüsse beschrieben werden sollen – beispielsweise als entweder saisonal oder intermittierend wasserführend.

Dies tritt in der praktischen Arbeit darin zu Tage, dass die OpenStreetMap-Community in der Vernetzung ihrer Daten darauf abzielt, primär und eventuell ausschließlich auf eine externe Datenhaltung in Wikidata aufzubauen. Die derzeitige Zielabsicht scheint zu sein, von OpenStreetMap-Datensätzen nur zu Wikidata-Datensätzen 1:1-Verknüpfungen herzustellen. Verbindungen zu anderen geografischen Datenbanken wie Pelagios oder GeoNames sollen offenbar überwiegend durch Datenkonsument:innen aus dem Datenbestand von Wikidata abgeleitet werden. Insbesondere die Verlinkung zu GeoNames bewegt sich derzeit im regulatorischen System der OpenStreetMap-Richtlinien in einer Grauzone faktischer Duldung.

Eignung der offenen Plattformen für DH-Arbeit

In der vorangegangenen Schilderung wurden bereits individuelle Charakteristika von Wikisource und OpenStreetMap dargestellt. Zu resümieren bleiben einerseits ihre Gemeinsamkeiten in der Verwendung als führende Datenspeicher im DH-Kontext – aber auch, inwiefern sie kontrastieren.

Die Praxiserfahrung über Jahre im vorgestellten Anwendungsfall bestätigt ausreichende Spielräume in der Datenmodellierung und -erfassung in beiden Systemen, sofern man ein Maß an pragmatischer

Kompromissbereitschaft einbringt. Wikisource ermöglicht primär eine layoutorientierte Texterfassung mit geringer semantischer Annotation, dafür aber grundsätzlich deutlichen Spielraum zur diplomatischen Transkription. Ähnlich verhält es sich in OpenStreetMap, das sich für die hier gesetzten Ziele als ausreichend flexibel zur Aufnahme der gewünschten Daten erwiesen hat. Bemerkenswert scheint allerdings die Positionierung der Community zu linked data – die sich hier im konkreten Anwendungsfall durch die normative Kraft des Faktischen reguliert hat.

Dieser Fall hebt hervor, dass beide Plattformen offene Datenspeicher mit Konsens- und Community-getriebener Regulierung sind. Sie bieten keine grundsätzliche oder verbürgte Hoheit oder vollständige Kontrolle über die eingearbeiteten Daten. Die beiden Projekte unterscheiden sich in diesem Zusammenhang in ihrer technischen Ausgestaltung, aber auch in der Art ihrer Daten:

Wikisource bietet sehr nützliche und ausgereifte Funktionalitäten zur Überwachung und Nachverfolgung von Veränderungen an Daten, die man im Blick behalten möchte. OpenStreetMap stellt keine äquivalenten Möglichkeiten zur Verfügung – was sich in der Praxis bereits als greifbares Manko bestätigt hat. Allerdings hat Wikisource den Vorteil, mit persistenten, diskreten, abgekapselt gelagerten Daten-Grundeinheiten (Einzelseiten) arbeiten zu können. Die Geo-Daten in OpenStreetMap hingegen sind hochvernetzt, was es konzeptuell deutlich erschwert, Daten-Einheiten zu isolieren und deren Veränderungen in einem Beobachtungs-Tool zu protokollieren.

Die APIs, die Wikisource und OpenStreetMap jeweils zur Datenabfrage bereitstellen, können sehr positiv bewertet werden. Funktionsumfang und Ausgabeformate beider APIs waren adäquat für alle im Referenzprojekt angestreten Ziele.

Offene Plattformen, Forschungsdateninfrastrukturen und research data engineering

Konzise angerissen aber nicht umfassend ergründet werden soll hier das Verhältnis von kostenfreien offenen Plattformen wie Wikisource oder OpenStreetMap zu in den DH betriebenen Datenspeichern, Editionsumgebungen und Forschungsdateninfrastrukturen. Ziel soll hier unter der eingangs genannten Prämisse der Abwesenheit einer Projekt-Finanzierung nur sein, kontrastive Überlegungen anzustoßen.

Als globale Tendenz formuliert, scheint es der status quo, dass zur Einrichtung und zum Betrieb von in den DH typischen Systemen sowohl Personalmittel wie auch eine institutionelle Anbindung dringend notwendig – fast unumgänglich – sind. Im Vergleich werden die Plattformen wie OpenStreetMap und Wikisource dauerhaft betrieben und kostenfrei zur Verfügung gestellt. Andererseits bieten die offenen Systeme nicht die Datensouveränität und die maßgeschneiderte Datenmodellierung, die in DH-Projekten erzielt werden können.

Grundsätzlich reagieren die DH auf dieses Spannungsfeld bereits mit dem wachsenden Fokus auf generischere Entwicklungen im RSE und durch den Aufbau von übergreifenden Forschungsdateninfrastrukturen. Es wird in der Zukunft eine pertinente wissenschaftshistorische Frage sein, rückblickend den gesteuerten Aufbau von Forschungsdateninfrastrukturen und das organische Aufgreifen von Wikidata in den DH zu vergleichen.

Die Erfahrungen im hier vorgestellten konkreten Referenzprojekt führen zu der Sicht, dass offene Plattformen eine gute Option für eine Gründungs- und Anwachsphase ohne Forschungsförderung sein können. Im hier vorliegenden inhaltlichen Kontext kann besonders hervorgehoben werden, dass das für marginalisierte Themen oder Forschende hochrelevant sein kann. Diese geschilderten Möglichkeiten negieren nicht die Vorteile eines finanzierten DH-Projekts. Vielmehr scheint die Perspektive vielversprechend, einen Projektstart auf offenen Plattformen als mögliche vorbereitende

Phase vor einer Förderung zu betrachten. Plattformen wie Wikisource und OpenStreetMap können in einem derartigen Szenario als ›Inkubatoren‹ für Forschungsdaten betrachtet werden.

Literatur

- Chatelard, Antoine. 1995. „Charles de Foucauld linguiste ou le savant malgré lui“. Études et Documents berbères 13 (2): 145–77.
- Foucauld, Charles de. 1940. Dictionnaire abrégé touareg-français de noms propres. Dialecte de l’Ăhaggar. Herausgegeben von André Basset. Larose Éditeurs. https://fr.wikisource.org/wiki/Dictionnaire_abrégé_touareg-français_de_noms_propres.
- Foucauld, Charles de. 1951. Dictionnaire touareg – français. Dialecte de l’Ăhaggar. Herausgegeben von André Basset. Imprimerie nationale de France. https://fr.wikisource.org/wiki/Dictionnaire_touareg_–_français.
- Ritter, Hans. 2009. Wörterbuch zur Sprache und Kultur der Twareg. Harrassowitz.