Critical AI in der Digitalen Editorik

Ries, Thorsten
thorsten.ries@austin.utexas.edu
University of Texas at Austin, USA
ORCID: 0009-0004-1112-8410

Andrews, Tara
tara.andrews@univie.ac.at
Universität Wien, Österreich
ORCID: 0000-0001-6930-3470

Rosendahl, Lisa
Lisa.Rosendahl@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland
ORCID: 0000-0002-4826-4553

Sahle, Patrick
sahle@uni-wuppertal.de
Bergische Universität Wuppertal, Deutschland
ORCID: 0000-0002-8648-2033

Viehhauser, Gabriel
gabriel.viehhauser@univie.ac.at
Universität Wien, Österreich
ORCID: 0000-0001-6372-0337

Vogeler, Georg
georg.vogeler@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0002-1726-1712

Hegel, Philipp
Philipp.Hegel@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland
ORCID: 0000-0001-6867-1511

Einleitung

Die wissenschaftliche digitale Edition war seit den Anfängen eine Triebfeder der Digital Humanities. Die frühesten Anwendungen digitaler Technologien wie etwa Robinsons *Canterbury Tales* und Busas *Index Thomisticus*, die digitale Edition von Musils Nachlass auf CD-ROM und DVD-ROM und die Entwicklung von TUSTEP, HNML und TEI waren editionswissenschaftliche Projekte, und die nachhaltige, skalierende wissenschaftliche Dokumentation und Erschließung von historischen Quellen, komplexen Textbefunden und kulturellem Erbe in Formaten wie Musik, Film, und born-digital Archiven gehört bis heute zu den wichtigsten Aufgaben der digitalen Geisteswissenschaften. Die Forschung zu Implementierung und Formaten digitaler wissenschaftlicher Edition hat sich in der Vergangenheit vor allem mit den praktischen und theoretischen Aspekten der Konzeption als Fortsetzung der historisch-kritischen Ausgabe, Datenmodellierung, Nachhaltigkeit, semantischer Erschließung und datentechnischer Anschlussfähigkeit gewidmet. Mit der gegenwärtigen Entwicklung von KI-Technologien, Machine Learning, LLM und multimodalen Transformern ergeben sich nicht nur neue Möglichkeiten, sondern auch Herausforderungen für die Theorie, Praxis, Konzeption, Aufgabenstellung und Rezeption digitaler Editionen. Auf verschiedenen Ebenen der Erstellung von digitalen Editionen hat KI bereits Einzug in den Workflow-Alltag gehalten (NER, semantische Erschließung, HTR, z.B. Wittgenstein Edition, Dehmel Digital, Schuchardt, Zweig usw., vgl. Pichler 2003, Pollin 2024, Pollin, Steiner und Zach 2023), auf anderen sind teils grundlegende Fragen ungeklärt (z.B. XAI, Reproduzierbarkeit, Fehlerquellen, Training-Bias und FAIRness, Urheberrechtsfragen, ökologische und ethische Fragestellungen, Attribuierbarkeit und Nachhaltigkeit von KI). In jüngerer Zeit hat sich „Critical AI“ als Forschungsfeld innerhalb der Digital Humanities etabliert, unter anderem mit einer wissenschaftlichen Zeitschrift unter dem Titel *Critical AI*. Das vorgeschlagene Panel versammelt Spezialist:innen für digitale Editionen, um in einen Austausch über die Chancen, Entwicklungen und Herausforderungen von KI in der digitalen Edition und Editionswissenschaft einzutreten. Es wird von der DH-Kommission der Arbeitsgemeinschaft für germanistische Edition organisiert (Koordination: Philipp Hegel, Thorsten Ries). Die Veranstaltung ist in vier Fragenkomplexe unterteilt:

1. Welche Rolle können Künstliche Intelligenzen im editorischen Arbeitsablauf spielen? Bei welchen Schritten können sie assistieren, welche Schritte übernehmen?
2. Wie verändert das die Rolle von Editor:innen und das Verständnis von Editionen? Wer wird in Zukunft für den Einsatz von Künstlicher Intelligenz in Editionsprojekten zuständig sein? Wie verändert sich das Verhältnis der Edition zu anderen wissenschaftlichen Aufgaben wie der Sprach- und Textanalyse?
3. Welche Probleme und Folgen sind zu erwarten und wie fehlertolerant dürfen wissenschaftliche Editionen sein? Wie reproduzierbar und nachhaltig sind die Ergebnisse?
4. Was kann und soll *Critical AI* im Hinblick auf digitale Editionen heißen?

Auf je zwei kurze Statements folgen in den ersten drei Fragekomplexen etwa fünfzehnminütige Diskussionen. Der vierte Komplex bündelt die Erkenntnisse aus den Impulsen und Diskussionen.

Veränderungen des editorischen Arbeitsprozesses

Tara Andrews

In den letzten Jahren konzentrierte sich die Diskussion über KI und ihre Möglichkeiten überwiegend auf generative KI – ihre Fähigkeit, oberflächlich betrachtet eine kohärente Antwort auf jede Frage geben zu können. Dies hat sowohl Begeisterung als auch Panik unter Philolog*innen hervorgerufen, die vorhersagen, dass ein entsprechend trainiertes Modell bald in der Lage sein wird, ganze kritische Editionen zu erstellen. Das Haupthindernis für diese Vision ist die Tatsache, dass jede angeforderte „kritische Interpretation" eines Textes in den meisten Fällen oberflächlich oder sogar halluzinatorisch wird. Künstliche Intelligenz geht jedoch weit über generative Methoden hinaus. Aktuelle Methoden des maschinellen Lernens im Allgemeinen – von HTR für die Transkription von Manuskripten bis hin zu Deep Learning für verschiedene Arten der Erkennung von literarischen Mustern – sind äußerst hilfreiche Entwicklungen für Philolog*innen. Hier gibt es sogar noch mehr Potenzial, das es zu erforschen gilt: Überall dort, wo die Interpretation, Encoding oder sogar stemmatische Analyse vorhersehbar und regelmäßig ist, kann uns KI in anderen Formen offensichtlich helfen. Trotz der aktuellen Fokussierung der Diskussion auf LLMs und generative KI sollte nicht übersehen werden, dass wir, so meine These, die Möglichkeiten anderer Formen der KI, insbesondere im Bereich der Knowledge Representation und des Reasonings, zur Unterstützung unserer Arbeit an Editionen noch (lange) nicht ausgeschöpft haben.

Gabriel Viehhauser

Die Frage, ob KI-Methoden über ihren Werkzeugcharakter hinaus zu konzeptionellen Umbrüchen in der Editorik führen, scheint sich je nach Fall unterschiedlich beantworten zu lassen: So eröffnen HTR-Verfahren die Möglichkeit, große Textmengen zu digitalisieren und damit durch das Druckzeitalter geprägte Vorstellungen von Kanon und philologischer Genauigkeit in Frage zu stellen. Demgegenüber erscheinen die Versuche von Entitäten- und Strukturerkennung mittels LLMs noch konventionell: Dass stochastische Methoden zu Unschärfe führen und einer kritischen Reflexion bedürfen, gilt für herkömmliche algorithmische Methoden auch; dass das editorische Geschäft notwendigerweise mit Unschärfen verbunden ist, ist schon in ‚traditionellen' Diskussionen bemerkt worden, etwa von Karl Stackmann, der bereits 1964 davon gesprochen hat, dass es zu den Aufgaben moderner Editionen gehöre, ein Höchstmaß an Unsicherheit über den Text auszustellen (Stackmann 1997, 23). In der Editorik hat das Aufkommen von digitalen Methoden dementsprechend gerade einem poststrukturalistisch geprägten Paradigmenwechsel zum Durchbruch verholfen, der von statischen, autorzentrierten hin zu dynamischen, überlieferungsgeschichtlich ausgerichteten Editionen geführt hat, die sich besonders für algorithmische Auswertungen anbieten. Dieser methodisch-theoretische Gleichklang scheint sich in den letzten Jahrzehnten verlaufen zu haben. Daran knüpfen sich zwei Fragen, nämlich, ob der Einsatz von KI zu einer ‚zweiten Renaissance' bzw. letztlich zu einer endgültigen Akzeptanz algorithmischer Methoden in der Editorik führen kann, und ob er die Felder Editorik und Textanalyse wieder zusammenbringen kann, indem Sprachmodelle stärker als bisher auf historische Sprachstufen und überlieferungsgeschichtliche Vielfalt Rücksicht nehmen.

Veränderung der Rolle von Editor:innen

Thorsten Ries

Das Potenzial von „intelligenten Editionen", bei denen KI eine halbautonome oder sogar autonome Rolle entweder im Produktionsprozess oder in der Nutzer:innen-Interaktion der Edition spielen würde, bedeutet einen Wandel in der Konzeption und technologischen Grundlage wissenschaftlicher Edition und der Rolle der Editor:in. KI kann bereits heute in editorischen Workflows eine Rolle spielen bei Alignierung von Texten und Varianten, HTR-basierten Basistranskriptionen und Aufgaben wie Fehlerkorrektur und Code-Erzeugung. Ein konsequenter Gedanke wäre, den von wissenschaftlichem Personal verfassten Kommentar vollständig durch ein fine-tuned LLM zu ersetzen, welches auf allen vorherigen Kommentaren und der neuesten Forschungsliteratur trainiert wurde und in einer digitalen Edition auf Anfrage Kontext- und Forschungsinformationen liefert. Das Problem der Autoritätsposition des wissenschaftlichen Kommentars könnte gelöst werden durch Leser:innen-Prompts, welche Fokus und Forschungsperspektive bestimmen. Die Editor:in wird immer noch einen Teil der philologischen Arbeit leisten und bestimmte Forschungsprobleme lösen, ihre Rolle wird sich jedoch in Richtung der Qualitäts- und Konsistenzkontrolle des Trainingsinputs und des Output verlagern. Ein solcher Paradigmenwechsel ist nicht ohne Schwierigkeiten: Reproduzierbarkeit, Kontrolle des Lernprozesses und Nichtdeterminierbarkeit sind bekannte Probleme, ferner Fragen der fairen Repräsentation innerhalb des Models und von Seiten eines LLM eingeführter Verzerrungen, Urheberrechts- und Datenschutzfragen (bei born-digital Materialien) sowie Fragen der Model-Nachhaltigkeit, Zitierbarkeit und Zurechenbarkeit wissenschaftlicher Leistungen.

Georg Vogeler

Mit der generativen KI bekommt das Verhältnis von Regelhaftigkeit und situativer Praxis editorischer Arbeit

eine neue Wendung: Editionswissenschaft versuchte, allgemeine Beobachtungen zu machen und Regelwerke zu erstellen, um die dokumentarische Textüberlieferung in die Gegenwart zu holen. Digitales Edieren zielte darauf, editorische Befunde in expliziten Datenmodellen auszudrücken und editorische Prozesse algorithmisch abzubilden. Mit maschinellem Lernen, z.B. HTR, lernen wir, die situativen Entscheidungen der Editor:innen an einen „Metaalgorithmus“ zu delegieren, der sich anhand von Trainingsmaterial dem editorischen Verhalten annähert. Generative KI hat nun zwei neue Dinge eingeführt: Erstens die Möglichkeit, den Output der Maschine als natürliche Sprache (oder Musik oder Bilder oder Videos) zu fassen, und zweitens die Integration von umfassendem Kontextwissen auf Seiten der Maschine. In vielen Experimenten lernen wir, dass generische Sprachmodelle auf Modelle editorischer Arbeit in Form z.B. der TEI zugreifen können und die Fähigkeit besitzen, Wissensbasen determinativ anzusprechen und selbständig einfache algorithmische Aufgaben zu lösen (Pollin, Steiner und Zach 2023, Czmiel u.a. 2024). In ihrer Funktion als Werkzeug lernen wir aber auch, dass sie in die Situationen eingebunden werden müssen, in denen bei dem:r Editor:in Spezialwissen bestimmter Hände, inhaltlicher Zusammenhänge oder literarischer Bezüge entsteht. Es scheint sich also eine Arbeitsteilung zwischen dem „Normalen“ und dem „Speziellen“ zu ergeben. Gleichzeitig müssen wir neue Formen des Umgangs mit Unzuverlässigkeit als Teil unserer editorischen Epistemologie erlernen, die damit neben das Talent, neben die Vertrautheit mit Sprache, Kultur und Text und neben die Vorsicht gegenüber der Verführungskunst der Texte treten, die Huygens im Jahr 2000 als Eigenschaften des:r Editor:in als Gegenpol zur Regelhaftigkeit der Editionspraxis benannt hat.

Erwartete Probleme und Folgen

Lisa Rosendahl

Das Schlagwort der Nachhaltigkeit steht im Zentrum der Bemühungen, kulturelle Artefakte mittels digitaler Editionen zu erhalten und verfügbar zu machen. Doch was nützen uns die FAIR-Prinzipien, wenn der Klimawandel ungebremst voranschreitet und neben katastrophalen Folgen wie der Zerstörung von Lebensräumen auch Serverräume überflutet werden oder längere Stromausfälle zu Datenverlusten führen? Bei der Planung, Erstellung und Archivierung digitaler Editionen sollten daher Maßnahmen zur Reduzierung des ökologischen Fußabdrucks berücksichtigt werden. Dies gilt in besonderem Maße für den Einsatz von KI, für deren enormen Ressourcenverbrauch in den Digital Humanities noch zu wenig Bewusstsein herrscht. Bei der Diskussion über die Nutzung von KI kann und darf dieser wichtige Aspekt neben all den vielversprechenden Vorteilen daher nicht außer Acht gelassen werden, wenn es uns um eine wirklich nachhaltige Bewahrung von Editionsgegenständen geht.

Patrick Sahle

Wir kennen die Prozesse und Anforderungen im editorischen Geschäft und wissen, wie wir von A (Überlieferung) nach B (Editionen) kommen wollen. Aktuelle KI-Anwendungen sind Werkzeuge, die darauf hin zu untersuchen sind, an welchen Stellen im Editionsprozess sie in welcher Weise produktiv eingesetzt werden können. Die aktuellen generativen KIs auf der Basis von LLMs haben spezifische Stärken und Schwächen, die einen Einsatz in bestimmten Bereichen eher nahelegen als in anderen. Auf dieser Basis müssen wir jetzt explorative Nutzungen für Anwendungsszenarien entwickeln und evaluieren. Dabei reichen die Optionen von einer direkten Nutzung generischer Anwendungen (wie ChatGPT) über das Entwickeln(-lassen) von Programmcode bis hin zu problemspezifischen Anwendungen und möglicherweise eigens erweiterten und trainierten LLMs. Dieser Blick in den Werkzeugkasten greift aber immer noch zu kurz, wenn er die Frage nach dem möglichen disruptiven Charakter der kommenden KI-Anwendungen unterschlägt. Es ist zu erwarten, dass diese nämlich nicht nur Werkzeuge für die gleichen Probleme sein, sondern unsere Verfahrensweisen, Fragestellungen und Perspektiven verändern werden. Zu den möglichen epistemologischen Folgefragen könnten solche gehören (1.) nach unserem Werkzeug-Begriff, (2.) dem Verhältnis zwischen dem Unschärfe-Paradigma der aktuellen KI-Anwendungen und dem auf Explizierung, Modellierung und Formalisierung beruhenden Kernprogramm der DH oder (3.) der Rolle kritischer Editionen im Wissenschaftssystem.

Critical AI

Das Panel versammelt Expert:innen aus dem Bereich der digitalen Editorik, die sich auf den Austausch mit Expert:innen aus dem Bereich der Künstlichen Intelligenz und mit Expert:innen aus anderen Bereichen der Digitalen Geisteswissenschaften, auch jenseits der Forschung an Texten, mit Interesse an diesen Entwicklungen freuen. Das Ziel des Austausches ist ein besseres Verständnis dessen, was Critical AI im Bereich der digitalen Editorik heißen soll.

Bibliographie

Beshero-Bondar, Elisa E. 2023. “Declarative Markup in the Time of ‘AI’: Controlling the Semantics of Tokenized Strings.” *Proceedings of Balisage* 28, Washington, DC. https://doi.org/10.4242/BalisageVol28.Beshero-Bondar01 (zugegriffen 21. November 2023).

Critical AI. Journal of Artificial Intelligence Studies . Duke University Press. Accessed July 22, 2024. https://www.dukeupress.edu/critical-ai (zugegriffen 22. Juli 2024).

Crymble, Adam. 2001. *Technology and the Historian: Transformations in the Digital Age* . Urbana: University of Illinois Press.

Czmiel, Alexander, Stefan Dumont, Franz Fischer, Christopher Pollin, Patrick Sahle, Torsten Schaßan, Martina Scholger, Georg Vogeler, Torsten Roeder, Christiane Fritze und Ulrike Henny-Krahmer. 2024. *Generative KI, LLMs und GPT bei digitalen Editionen* . Workshop bei *DH quo vadis* (DHd 2024), Passau, 27. Februar 2024. https://drive.google.com/drive/u/0/folders/1z_j1awaX4gRUU_Ykb1TqUhzoHQLnZXed (zugegriffen 23. Juli 2024).

Dehmel_digital. hg. von Julia Nantke. Hamburg: Hamburg University. https://dehmel-digital.de/ (zugegriffen 22. Juli 2024).

Dobson, James E. 2023. "On Reading and Interpreting Black Box Deep Neural Networks". In *International Journal of Digitals Humanities* 5, 431–449. https://doi.org/10.1007/s42803-023-00075-w

El-Hajj, Hassan, Oliver Eberle, Anika Merklein, Anna Siebold, Noga Schlomi, Jochen Büttner, Julius Martinetz, Klaus-Robert Müller, Grégoire Montavon und Matteo Valleriani. 2023. "Explainability and Transparency in the Realm of Digital Humanities: Toward a Historian XAI". In *International Journal of Digital Humanities* 5, 299–331. https://doi.org/10.1007/s42803-023-00070-1

Fusi, Daniele. 2009. "Aspects of Application of Neural Recognition to Digital Editions" . In: *Kodikologie und Paläographie im digitalen Zeitalter* , hg. v. Malte Rehbein, Patrick Sahle und Torsten Schaßen, 175–195. Norderstedt: Books on Demand.

Graziosi, Barbara, Johannes Haubold, Charlie Cowen-Breen und Creston Brooks. 2023. "Machine Learning and the Future of Philology: A Case Study." In *TAPA* , 153, Nr. 253–284. https://doi.org/10.1353/apa.2023.a901022 (zugegriffen 24. Juli 2024).

Huygens, R.B.C. 2000. *Ars edendi. A Practical Introduction to Editing of Medieval Latin Texts* . Turnhout: Brépols.

Pichler, Alois. 2023. "Interactive Dynamic Presentation (IDP) and Semantic Faceted Search and Browsing (SFB) of the Wittgenstein Nachlass." In *Wittgenstein-Studien* 14, Nr. 1, 131–151.

Pollin, Christopher. 2024. „AI-Datenerzeugung in der digitalen Briefedition." Präsentation bei *Zwischen Tinte und Code. Zu Stefan Zweigs Briefen im Datenzeitalter* . Salzburg, 5. Februar 2024. https://docs.google.com/presentation/d/1as04Dbas-1cCfTqdNJEkA8a_cYoVac7uhnGSercffYM/edit?usp=sharing (zugegriffen 22. Juli 2024).

Pollin, Christopher, Christian Steiner und Constantin Zach. 2023. "New Ways of Creating Research Data: Conversion of Unstructured Text to TEI XML using GPT on the Correspondence of Hugo Schuchardt with a Web Prototype for Prompt Engineering." Präsentation bei *Anything goes?* (FORGE 2023), Tübingen, 6. Oktober 2023. https://docs.google.com/presentation/d/1wilgLV1mm8xria4yvaUkghiO7G534JyxRjjR45RkMWw/edit?usp=sharing (zugegriffen am 22. Juli 2024).

Pollin, Christopher, Martina Scholger, Elisabeth Steiner, Sarah Lang, Selina Galka und Sebastian Schiller-Stoff. 2024. " Project Overhaul und Refactoring der digitalen Edition der 'Urfehdebücher der Stadt Basel' mithilfe von GPT-4 und LLM." Präsentation bei *DH quo vadis* (DHd2024). Passau, 5. Februar 2024. https://docs.google.com/presentation/d/1OTasxB8FP9-n1-ghr3qtX7V8bf-tBGXDDIfbla5FwSk/edit?usp=sharing (zugegriffen 22. Juli 2024).

Ries, Thorsten, Karina van Dalen-Oskam und Fabian Offert. 2024. "Reproducibility and Explainability in Digital Humanities." In *International Journal of Digital Humanities* 6, 1–7. https://doi.org/10.1007/s42803-023-00083-w

Stackmann, Karl. 1997. *Mittelalterliche Texte als Aufgabe. Kleine Schriften 1* , hg. von Jens Haustein, Göttingen: Vandenhoeck und Ruprecht.

Stiemer, Haimo, Evelyn Gius und Dominik Gerstorfer. 2024. "Künstliche Intelligenz und literaturwissenschaftliche Expertise." In *KI:Text. Diskurse über KI-Textgeneratoren* , hg. von Gerhard Schreiber und Lukas Ohly, 455–466. Berlin, Boston: De Gruyter.

Stutzmann, Dominique. 2023. Automatische Texterkennung (ATR) in der Mediävistik: Werkstattbericht zu den Projekten Himanis und Home und neue Perspektiven für Historiker:innen. Präsentation beim „ *Digital History"-Forschungskolloquium* , Berlin, 24. Juni 2023.