Tim Geelhaar

HAMSTERRAD ODER HIMMELSLEITER? ODER WARUM DIE DIGITALISIERUNG SO ENDLOS SCHEINT

Wer kennt sie nicht, die großen Ziele, die hehren Ansprüche, die hohen Erwartungen? Wie überall gilt auch in der Wissenschaft die Digitalisierung als Hoffnungsbringer – wenn auch nicht mehr als Heilsbringer. Fast schon hat es den Anschein einer ausgebliebenen Parusie. Ist die Digitalisierung doch nicht unsere Himmelsleiter in das Paradies der Forschung? Stecken wir fest im Hamsterrad des schier endlos scheinenden Preprocessing? Gut 20 Jahre nun wird unter dem Label ›Digital Humanities‹ an der digitalen Transformation der Geisteswissenschaften gearbeitet; Visionen, Experimente und Projekte gibt es schon erheblich länger.[1] Und trotz aller Errungenschaften scheint es so, als würden viele Versprechen aufgestellt, aber nur wenige eingelöst. Was nützen *proof-of-concept*s, wenn die damit beworbenen Arbeitstechniken oder Datenbestände keine Anwendung finden, weil diese unausgereift bleiben? Wie viel Arbeit nimmt uns die Digitalisierung ab, wenn wir immer wieder Daten zusammensuchen und aufbereiten müssen und sich die Analyse immer weiter nach hinten verschiebt? Reproduzieren wir nicht ähnliche Rezeptionsphänomene, die wir aus der Zeit analoger Bibliotheken kennen? Große Säle voller vor sich hin einstaubender Buchbestände gleichen der Vielzahl an digitalen Editionen, die nach einigen Jahren der Existenz ohne Betreuung im Niemandsland des Digitalen verloren gehen.[2] Die Kanonbildung durch ständige Rezeption weniger ausgewählter Werke hingegen entspricht der Wiederverwendung der immer gleichen Datensätze – wie der *Patrologia Latina (PL)*[3] auf den Gebieten der Geschichte und Theologie.

1 Vgl. Manfred Thaller: I Grundlagen – Geschichte der Digital Humanities, in: Digital Humanities – eine Einführung, hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein, Stuttgart 2017, S. 3-12; Isabelle Guyot-Bachy: Conclusions, in: Digitizing Medieval Sources – L'édition en ligne de documents d'archives médiévaux. Challenges and Methodologies – Enjeux, méthodologie et défis, hg. von Christelle Balouzat-Loubet ARTEM 27), Turnhout 2019, S. 173-178.

2 Man siehe nur die Anzahl an Editionen im Catalogue Digital Editions von Greta Franzini (https://dig-ed-cat.acdh.oeaw.ac.at/, Zugriff: 3. Mai 2023).

3 Ursprünglich in 222 Druckbänden herausgegebene Sammlung von kirchlich rele-

DOI https://doi.org/10.15499/kds-005-016 | CC BY-SA 4.0

Diese durchaus polemische Zuspitzung beschreibt eine persönliche Erfahrung. Sie deutet an, warum es dem Autor wichtig ist, über Erfahrungen und Erwartungen im Zuge der Digitalisierung zu reflektieren. Denn diese können zuweilen stark divergieren. Erst jüngst hat die Einführung von ChatGPT[4] dem Gedankenkarussell um Hoffnungen, Wünsche, Skepsis, Ängste und Ablehnung neuen Anschub verliehen. Andererseits wissen wir nur zu gut, wie sich scheinbar endlos all die Aufbereitungsschritte von Daten hinziehen können, bevor man überhaupt zu deren inhaltlichen Aus- und Bewertung kommt. Einen Goldstandard oder »smart data«[5] zu erhalten, kann das eigentliche Ziel, digitale Sammlungen auf computergestützte Weise auszuwerten, in weite Ferne rücken lassen. Daher scheint es umso dringlicher darüber nachzudenken, woher unsere Erwartungen kommen, warum sie sich öfter nicht erfüllen lassen, welche Konsequenzen sich aus der Nichterfüllung ergeben und wie ein angepasstes Erwartungsmanagement negative Reaktionen vermeiden hilft. Schließlich gilt es zu diskutieren, wie die intensive und hochproduktive Arbeit im Bereich der Digital Humanities eine breitere Wertschätzung und Aufnahme erfährt.

Um diesen Fragen nachzugehen, bietet der erste Teil einen skizzenhaften Überblick über die mehr als 18 Jahre andauernde Arbeit an einer analytischen Volltextdatenbank lateinischer Texte des Mittelalters. Dieses Projekt basiert auf der Zusammenarbeit von Geschichtswissenschaft und Computerlinguistik, es verbindet Ansätze der Historischen Semantik mit Methoden der Corpuslinguistik und des Text Mining. Der zweite Teil widmet sich den in diesem Projekt gewonnenen Erfahrungen,[6] bevor Überlegungen zum Er-

vanten Texten von Tertullian (um 200 nach Christus) bis Papst Innozenz III. (gest. 1216), siehe Patrologiae cursus completus. Series Latina, hg. von Jacques-Paul Migne, Paris 1844-1865. Die PL wurde 1993 vom britischen Verlag Chadwyck-Healey retro-digitalisiert, als CD-Rom und ab 1997 als kostenpflichtige Datenbank im Netz angeboten. Jetzt vertreibt der amerikanische Verlag Proquest die kommerzielle Fassung, ohne aber die Funktionalitäten an die Bedürfnisse des Text Mining angepasst zu haben. Es existieren aber diverse Digitalisate und digitale Volltexte im Netz, zum Beispiel auf Corpus Corporum (https://mlat.uzh.ch/, Zugriff: 3. Mai 2023).

4 Chatbot des Generative Pre-trained Transformers (OpenIA), veröffentlicht zum Testen am 30. November 2022 auf der Webseite https://openai.com/blog/chatgpt/ (Zugriff: 3. Mai 2023).

5 Vgl. Christof Schöch: Big? Smart? Clean? Messy? Data in the Humanities, in: Journal of Digital Humanities 2(3), 2013 (http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/, Zugriff: 3. Mai 2023).

6 Der Autor dieses Beitrags arbeitet seit 2013 an diesem Projekt in den Bereichen Sammlungs- und Corpusaufbau, Daten- und Metadatenmodellierung, Vorverarbeitung, Lemmatisierung, Postlemmatisierung, Lexikonpflege und Vermittlung dieser

wartungsmanagement den Beitrag beschließen. Besonders hervorzuheben ist generell ein noch nicht hinreichend beachteter Faktor für die Langwierigkeit solcher Großprojekte: die unterschiedlichen Geschwindigkeiten verschiedener Prozesse. Dies betrifft einerseits den Prozess der interdisziplinären Projektarbeit selbst, aber weit mehr noch den Prozess der Rezeption und Adaption neuer digitaler Techniken und Methoden in den Geisteswissenschaften. Letzteres kann sich aus strukturellen Gründen Jahre hinziehen – Jahre, in denen die zu adaptierenden Techniken bereits wieder überholt oder gar aufgegeben worden sein können. Diese Zeitverschiebungen sind schwerlich zu vermeiden. Doch sollten wir uns darum bemühen, Wege des Umgangs mit diesem Phänomen zu finden, gerade damit sich die Digitalisierung nicht endlos in die Länge zieht.

1. Der lange Weg zum Latin Text Archive

Bereits Mitte der 2000er Jahre haben der Mediävist Bernhard Jussen und der Computerlinguist Alexander Mehler damit begonnen, ein Managementsystem für eine corpusbasierte historische Semantik zu entwickeln. Das schon 2007 publizierte Ziel lautete, chronologisch organisierte Corpora lateinischer Texte auf vom Computer erfassbare, synchrone wie diachrone Sprachwandelphänomene zu untersuchen, um so semantischen und daraus abgeleitet historischen Wandel erkennen zu können.[7] Etwa zeitgleich ist Bernhard Jussen mit dem Gottfried Wilhelm Leibniz-Preis ausgezeichnet worden. Das Preisgeld bot daraufhin die Chance, die gemeinsamen Pläne auf Alexander Mehlers webbasierter Arbeitsplattform, dem *eHumanities-Desktop*,[8] umzusetzen.

Datenbank. Die behandelten Erfahrungen beschränken sich auf die mediävistische Perspektive des Autors. Erfahrungen aus technologischer Perspektive müssen daher an dieser Stelle ausgeklammert bleiben.

7 Vgl. Bernhard Jussen, Alexander Mehler und Alexandra Ernst: A Corpus Management System for Historical Semantics, in: Sprache und Datenverarbeitung, in: International Journal for Language Data Processing 31(1-2), 2007, S. 81-89; hier S. 81f. Vgl. auch Alexander Mehler u.a.: Der eHumanities Desktop als Werkzeug in der historischen Semantik: Funktionsspektrum und Einsatzszenarien, in: JLCL 26(1), 2011, S. 97-117; hier S. 101. Alle genannten Aufsätze von und mit Alexander Mehler sind auf der Webseite des Texttechnology Lab www.hucompute.org (Zugriff: 3. Mai 2023) als pdf erhältlich.

8 Vgl. https://hudesktop.hucompute.org/index.jsp (derzeit nicht erreichbar, Zugriff: 3. Mai 2023).

Das *Historical Semantic Corpus Management*, kurz HSCM, wurde dabei ein Modul innerhalb eines kompletten texttechnologischen Ökosystems: Das Administrationsmodul organisierte die Benutzer- und Rechteverwaltung für die kollaborative Arbeit. Der Corpus-Manager diente dem Aufbau und der Verwaltung von Textressourcen. Das Annotator-Modul ermöglichte die Auszeichnung von Textressourcen mit Metadaten. Der Preprocessor überführte Rohdaten nach TEI P5, tokenisierte, zeichnete Wortarten aus und lemmatisierte. Hierfür griff das System auf das eigens erstellte Wortformenlexikon zurück, das mittels des Lexikon-Browsers verwaltet wurde.[9] Der Reiz an diesem Ökosystem war, dass es verschiedensten Projekten erlaubte, mit ihren eigenen Materialien zu arbeiten, für alle natürlichen Sprachen gleichermaßen nutzbar und modular erweiterbar war.

Solch ein System entsteht nicht über Nacht. Zwischen 2008 und 2014 wurden sämtliche Module realisiert, wie auch ein Text Classifier und ein Wiki für Dokumentationszwecke. Sehr schnell waren Worthäufigkeitsstatistiken, Kookkurrenzanalysen über einen oder eine beliebige Anzahl von Texten möglich, Vergleiche von Ergebnissen waren ebenso machbar wie eine der Zeit entsprechende Visualisierung. Auch der Download von Konkordanztabellen und sonstiger Ergebnisse zur Weiterverarbeitung in anderen Systemen war gegeben. Jedoch kam es immer wieder zu neuen Herausforderungen. Das System erforderte von technischer Warte aus eine konstante Betreuung und immer wieder wurden Anpassungen an der Bedienungsoberfläche nötig, da die Bedienbarkeit der verschiedenen Module für viele Nutzer:innen zu kompliziert war. Die Lemmatisierung war nicht ausgereift und ließ sich nur in beschränktem Maße händisch korrigieren. Erschwerend hinzu kam, dass die anfängliche Textmenge größenmäßig und hinsichtlich ihrer Textsortenzusammensetzung für verlässliche Studien ungenügend war, obwohl das Gesamtkorpus aus 95 Millionen Wörtern bestand.[10] Schließlich war die mittels Information Retrieval gewonnene Metadatenannotation für geisteswissenschaftliche Zwecke unbrauchbar. Das lag vor allem an der Unzuverlässigkeit der gewonnenen Metadaten. Viele Texte waren in der Ausgangsedition falschen Autoren[11] und somit dem falschen Jahrhundert zugeordnet. Mit anderen Worten: Der Weg zu einer corpusbasierten, diachronen Langzeituntersuchung von Sprachwandelphänomenen über ein texttechnologisches Ökosystem war konzeptionell der richtige und wäre auch heute noch der zu

9 Vgl. Mehler u.a. eHumanities Desktop (Anm. 7).

10 Das erste Corpus bestand aus der PL (Anm. 3).

11 Da die PL eine Sammlung der Schriften von Kirchenvätern ist, wird hier bewusst nicht gegendert.

bevorzugende. Die Umsetzung jedoch verlief viel schleppender als von allen Seiten gewünscht. Viele Herausforderungen wurden erst im Verlauf sichtbar und die geisteswissenschaftliche Seite musste erst einmal lernen, welche Erfordernisse mit welchem Aufwand zu bewältigen waren.

In den Folgejahren konnten viele Herausforderungen in anderen Projektzusammenhängen angegangen werden. Zu nennen sind das LOEWE-Schwerpunktprogramm *Digital Humanities – Integrierte Aufbereitung und Auswertung textbasierter Corpora* in den Jahren 2011-2014 sowie das BMBF-Projekt *Computational Historical Semantics* von 2013 bis 2016.[12] In dieser Zeit wurden von informatischer Seite ein TextReuse-Modul, eine Netzwerkanalyse von Kookurrenzen sowie das Lemmatisierungstool *TTLAb Tagger* neu entwickelt, das die Zuverlässigkeit der automatischen Lemmatisierung deutlich erhöhte.[13] Indem der Tagger ins Preprocessing auf dem eHumanities-Desktop integriert wurde, verbesserten sich auch die Ergebnisse der Kollokationsanalysen. Außerdem wurde ein *Lemmatisation Editor* implementiert, der es Redakteur:innen erlaubte, jede einzelne Auszeichnung zu überprüfen und zu korrigieren. Dies ermöglichte gleichzeitig Anpassungen und Erweiterungen am *Frankfurt Latin Lexicon,*[14] dem bereits erwähnten Wortformenlexikon, was wiederum die Lemmatisierung aller Texte verbesserte. Im Gegensatz zu statischen korpusanalytischen Tools wie der *SketchEngine* oder der *CorpusWorkbench*[15] war der textuelle Gesamtbestand jederzeit an jeder Stelle auf XML-Ebene veränderbar, was Qualitätskontrolle und -steigerung ermöglichte.

12 Vgl. Integrierte Aufbereitung und Auswertung textbasierter Corpora (https://pro loewe.de/de/loewe-vorhaben/nach-themen/digital-humanities-hessen/) sowie https://comphistsem.org/ für das BMBF-Projekt (Zugriff: 4. Mai 2023).

13 Vgl. Steffen Eger, Rüdiger Gleim und Alexander Mehler: Lemmatization and morphological tagging in German and Latin: A Comparison and a survey of the state-of-the-art, in: Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016; Tim vor der Brück und Alexander Mehler: TLT-CRF: A Lexicon-supported Morphological Tagger for Latin Based on Conditional Random Fields, in: Proceedings of the 10th International Conference on Language Resources and Evaluation, 2016; Rüdiger Gleim, Alexander Mehler u.a.: A practitioner's view: a survey and comparison of lemmatization and morphological tagging in German and Latin, in: Journal of Language Modeling 7/1, 2019, S. 1-52.

14 Vgl. Alexander Mehler u.a.: The Frankfurt Latin Lexicon. From Morphological Expansion and Word Embeddings to SemioGraphs, in: Studi e Saggi Linguistici, 58/1, 2020, S. 121-155.

15 Vgl. SketchEngine (https://www.sketchengine.eu/) und IMS Open Corpus Workbench (https://cwb.sourceforge.io/) (Zugriff: 4. Mai 2023).

Kooperationsvereinbarungen halfen, neues Textmaterial zu integrieren. Insbesondere die *Monumenta Germaniae Historica* stellten die XML-Versionen einzelner Editionsbände aus ihrem openMGH-Projekt[16] zur Verfügung – was jedoch wieder zu erhöhtem Nachbearbeitungsaufwand führte, um die Texte mit dem *TTLab Tagger* zu lemmatisieren. Das Problem der Textklassifizierung wurde durch eine selbstentwickelte binomische Texttypenklassifikation gelöst, die zwischen Textart und Textfunktion unterscheidet und Mehrfachzuordnungen zulässt, jedoch nicht automatisiert Texte klassifiziert.[17]

Am wichtigsten in der Außendarstellung war die Entwicklung der Webplattform *Comphistsem.org*.[18] Diese sollte mit einem zeitgemäßen Web-Design den Einstieg in die Verwendung von HSCM erleichtern und seine Akzeptanz unter den Mediävist:innen durch eine einfachere Bedienbarkeit steigern. Und dies gelang: Das DFG finanzierte Lehnswesen-Projekt von Jürgen Dendorfer bereitete zum Beispiel die Kaiser- und Königsurkunden der Stauferzeit in HSCM auf und das Akademievorhaben zu den fränkischen Herrschererlassen von Karl Ubl lemmatisierte die sogenannten Kapitularien der Merowinger- und Karolingerzeit.[19] Im BMBF-Projekt *Humanist Computer Interaction auf dem Prüfstand* nutzte das Team um Marietta Horster HSCM zur Lemmatisierung von Cassiodors Variae.[20]

16 Vgl. Clemens Radl, Benedikt Marxreiter und Bernd Posselt: Die MGH im digitalen Zeitalter, in: Mittelalter lesbar machen: Festschrift 200 Jahre Monumenta Germaniae Historica; Grundlagen, Forschung, Mittelalter, hg. v. Martina Hartmann, Annette Marquard-Mois und Arno Mentzel-Reuters, Wiesbaden 2019, S. 39-53; hier S. 43. Vgl. auch Monumenta Germaniae Historica. Verfügbare Bände geordnet nach MGH-Abteilungen und Reihen openMGH (https://www.mgh.de/de/mgh-digital/openmgh/verfuegbare-baende, Zugriff: 4. Mai 2023).

17 Vgl. Latin Text Archive. Text Type Classification (https://lta.bbaw.de/d/text-type-classification, Zugriff: 4. Mai 2023).

18 Die Seite https://comphistsem.org/ (Zugriff: 4. Mai 2023) ist zwar noch erreichbar, aber da der eHumanities-Desktop zur Zeit nicht erreichbar ist, wird mittlerweile auf das LTA verwiesen. Zu comphistsem.org vgl. auch Roberta Cimino, Tim Geelhaar und Silke Schwandt: Digital Approaches to Historical Semantics: new research directions at Frankfurt University, in: Storicamente 2015, 1-16 DOI 10.12977/stor594.

19 Vgl. Die Formierung des Lehnswesens (https://www.lehnswesen.uni-freiburg.de/) und Capitularia – Edition der fränkischen Herrschererlasse (https://capitularia.uni-koeln.de/project/). Die Daten sind jetzt über https:/lta.bbaw.de abruf- und weiterverwendbar. Vgl. Bernhard Jussen und Karl Ubl (Hg.): Die Sprache des Rechts (Historische Semantik 33), Göttingen 2022. (Zugriff: 4. Mai 2023).

20 Vgl. Humanist Computer Interaction auf dem Prüfstand (https://humanist.hs-mainz.de/, Zugriff: 4. Mai 2023).

Auch nach Projektende ging die Arbeit über kleinere Projektfinanzierungen und durch persönlichen Einsatz weiter. Neue Texte mussten eingepflegt, lemmatisiert und korrigiert werden. Zudem wurde 2015 damit begonnen, die Texte präziser zu datieren und VIAF-Angaben für Autoren und Werke zu ergänzen. Dank neuer Kooperationen mit Textgebern kamen wir den gewünschten Korpora näher, konnten aber immer noch keine Referenzkorpora bereitstellen. Nur die Aufbereitung der Urkunden der Abtei Cluny erbrachte ein brauchbares Korpus.[21]

Immer drängender wurde das Problem der langfristigen Absicherung. Wie ließ sich *comphistsem.org* aufrechterhalten, zumal im so wechselhaften, universitären Umfeld? Sowohl die redaktionelle als auch die technische Betreuung konnte nur noch zu Lasten anderer Aufgaben bewältigt werden. Da Bernhard Jussen zwischenzeitlich Leiter des Mittelalterzentrums an der Berlin-Brandenburgischen Akademie der Wissenschaften geworden war, schien es der beste Weg, die Datenbank unter dem Namen *Latin Text Archive (LTA)* von der Universität an die Akademie zu überführen. Ein Prototyp konnte bereits auf der DHd-Tagung 2019 vorgestellt werden.[22]

Im gleichen Jahr präsentierte Alexander Mehler mit dem *Semiographen* ein Analysewerkzeug, mit dem sich Wordembeddings visualisieren lassen. Erste Testergebnisse mit den lemmatisierten lateinischen Textbeständen wurden 2020 publiziert.[23] Dies ist für die historisch-semantische Analyse hilfreich, weil semantische Bezüge sichtbar werden, die nicht den Bias menschlicher Vorannahmen durchlaufen haben. Eine zweite, sehr sinnvolle Erweiterung für den analytischen Werkzeugkasten war die Verknüpfung des LTA mit *Diacollo,*[24] das in der Lexikografie gern genutzte Tool zur Auswertung und Visualisierung von Kollokationen in diachron organisierten Korpora. Nach Corona-bedingten Verzögerungen wird nun eine Langzeitfinanzierung beantragt, um zum einen die langfristige institutionelle Absicherung zu erreichen und zum anderen den vollen Funktionsumfang des LTA zu realisieren. Vor allem sollen endlich die Bestände so erweitert werden, dass den Nutzenden ermöglicht würde, belastbare, nachvollziehbare und überprüfbare Ergebnisse

21 Vgl. LTA. Corpus of Cluny Charters (https://lta.bbaw.de/corpus/cluny, Zugriff: 4. Mai 2023).

22 Tim Geelhaar: Das Latin Text Archive (LTA) – Digitale Historische Semantik von der Projektentwicklung an der Universität zur Institutionalisierung an der Akademie, in: DHd 2019 Digital Humanities: multimedial & multimodal, Konferenzabstracts, hg. von Patrick Sahle, S. 266-268.

23 Vgl. Mehler, Frankfurt Latin Lexicon (Anm. 14).

24 Vgl. DiaCollo: Kollokationsanalyse in diachroner Perspektive (https://www.clarin-d.net/de/kollokationsanalyse-in-diachroner-perspektive, Zugriff: 4. Mai 2023).

aus der Korpusanalyse gewinnen, visualisieren und veröffentlichen, also gemäß der FAIR-Prinzipien nachnutzbar anbieten zu können. Auch im Jahr 2023 ist ein avanciertes Text Mining für das Lateinische auf Grundlage von historischen Corpora immer noch ein Desiderat.[25]

2. Erfahrungen aus der Projektarbeit

Dieser Schnelldurchlauf durch die Geschichte dieser Datenbank veranschaulicht, dass sich die Realisierung eines großen Ziels zuweilen scheinbar endlos in die Länge ziehen kann. Allerdings wurde die Datenbank aus dem Nichts geschaffen und wir haben bei ihrer Umsetzung vermutlich erst alle möglichen Erfahrungen sammeln müssen, die man für die Realisierung eines digitalen Mammutprojektes braucht.

Drei Erfahrungen möchte ich herausgreifen. Eine erste Hürde war der Zeit- und Arbeitsaufwand bei der Suche nach und der Umwandlung von Texten. Mittlerweile gibt es gute Lösungen, das heißt *ready-to-use*-Lösungen für Geisteswissenschaftler:innen, die Vorverarbeitungsaufgaben übernehmen. Die Bielefelder Plattform *nopaque* bietet einen einfachen Weg, um aus Fotos oder Scans digitale Volltextdateien zu erstellen und auszuwerten.[26] Doch selbst wenn Scan, OCR-Erkennung und NLP-Transformation vereinfacht werden, erfordert die Aufbereitung von digitalen Volltexten immer noch ein hohes Maß an Textverständnis, Zeit und Akribie. So mussten sich die Geisteswissenschaftler:innen mit den Grundlagen digitaler Editorik vertraut machen, um aus den 125 openMGH-Bänden die mehr als 6.000 Einzeltexte als einzelne TEI/XML-Dateien mit passender Textstruktur zu extrahieren. Dies war ursprünglich nicht geplant, zumal sich das LTA nicht als Editionsvorhaben versteht, sondern Texte für eine gemeinsame Analyse nach der Edition zusammenführt. Diese Kenntnisse helfen uns nun aber bei der Evaluierung und Überführung von digitalen Textbeständen unserer Kooperationspartner wie dem Albertus-Magnus-Institut.

25 Während Datenbanken wie ALIM (http://it.alim.unisi.it/) oder Corpus Corporum (https://www.mlat.uzh.ch/) zumindest eine Volltextsuche, aber kein Korpusmanagement [Corpus-Management, s.o.] anbieten, zeigen Projekte wie Cartae Europae Medii Aevi (CEMA) (https://cema.lamop.fr/) an, was künftig möglich werden sollte (Zugriff: 4. Mai 2023).

26 Vgl. Patrick Jentsch und Stephan Porada: From Text to Data. Digitization, Text Analysis and Corpus Linguistics, in: Silke Schwandt (Hg.): Digital Methods in the Humanities. Challenges, Ideas, Perspectives, Bielefeld 2020, S. 89-128. Der Service ist erreichbar unter https://nopaque.uni-bielefeld.de (Zugriff: 4. Mai 2023).

Zweitens stellte der Zeit- und Modellierungsaufwand bei der Metadatenannotation das Projekt vor Herausforderungen. Immer wieder müssen fachwissenschaftlich abgesicherte Informationen erst ermittelt werden. Außerdem müssen individuelle und zu dokumentierende Entscheidungen getroffen werden, zum Beispiel, wie schwierige Datierungen und Ortsangaben aufzunehmen sind. Verzögerungen treten hier auf, wenn sich die Ansprüche im Laufe der Zeit ändern und eine immer tiefere Erschließung erfordern. So kamen im Laufe der Zeit Verknüpfungen zu externen Ressourcen wie den *geschichtsquellen.de*[27] hinzu.

Drittens ist der Zeitaufwand bei der Lemmatisierung zu bedenken. Wie Menschen so können auch die besten Lemmatisierungstools nicht immer richtige Entscheidungen treffen. Nachlemmatisierung ist daher eine wichtige, wenngleich zeitraubende Angelegenheit. Zudem ist hier nur von der morphologischen und noch nicht von syntaktischer oder semantischer Auszeichnung die Rede. Für diese weiteren, notwendigen Ebenen fehlen bisher verlässliche Tools und Lexika. Das syntaktische Parsing wird für Latein immer noch entwickelt und erst seit 2019 gibt es überhaupt den Versuch, ein LatinWordNet nach Vorbild des WordNets zu erstellen,[28] das wir künftig mit unseren Beständen verknüpfen können.

3. Vom Umgang mit Erwartungen

In Erfahrungen spiegeln sich Erwartungen. Hier sind drei Perspektiven zu berücksichtigen: die Erwartungen des eigenen Teams, der Nachnutzenden und der Mittelgeber. Wie deutlich wurde, mussten sich die Geisteswissenschaftler:innen recht bald von der Erwartung verabschieden, schnell ein ausgereiftes und im Wissenschaftsdiskurs anschlussfähiges Arbeitsinstrument in Händen zu halten. Das interdisziplinäre Team musste zudem feststellen, dass Erwartungen schnell divergieren können. So hatte die informatische Seite ihre eigenen Ziele sehr viel früher erreicht als die geisteswissenschaftliche. Denn erst als das System produktionsfähig war, konnten die Geisteswissenschaftler:innen die Datenaufbereitung angehen, was auch heißt, dass die Phase der Datenanalyse noch weit entfernt war. Überdies ergab sich eine operative Schwierigkeit, denn für die Datenpopulation und

27 Dabei handelt es sich um das digitale Repertorium zu den Quellen des deutschen Mittelalters an der Bayerischen Akademie der Wissenschaften.

28 Vgl. https://latinwordnet.exeter.ac.uk/ für das Latin WordNet und für das englische WordNet https://wordnet.princeton.edu/ (Zugriff: 4. Mai 2023).

allfällige Änderungswünsche war immer wieder technischen Support vonnöten. Hier wird die Aufspaltung in *theoretical* und *applied* Digital Humanities sichtbar.[29] Denn ab einem bestimmten Zeitpunkt waren erst einmal nur angewandte Digital Humanities notwendig, während es für die Informatik keine Herausforderungen mehr gab. Solche Verschiebungen innerhalb fachübergreifender Gemeinschaftsarbeit sind bei digitalen Projekten immer möglich und zu berücksichtigen, um Enttäuschungen vorzubeugen.

Ein anderer Grund für enttäuschte Erwartungen ist sehr viel geläufiger. Ganz allgemein verbindet sich mit der Digitalisierung die Hoffnung, sehr viel mehr, sehr viel einfacher in sehr viel kürzerer Zeit zu erreichen und sogar Aufgaben zu bewältigen, die wir ohne sie nicht schaffen könnten.[30] Was meist übersehen wird, ist die manuelle Kärrnerarbeit, die in die Datenbanken einfließen muss, bevor sich der gewünschte Skalierungseffekt einstellt. Plakativ ließe sich sagen, Digitalisierung bereitet mehr Arbeit und nicht weniger. Das gilt nicht nur für die Datenpopulation, sondern auch für die Annotation, wie Experimente mit dem Annotationstool CATMA[31] oder der Datenbank HEURIST[32] gezeigt haben. Der Lohn besteht hingegen in der Transparenz im Umgang mit den Quellen, der Nachnutzbarkeit des Materials und den zusätzlichen Auswertungsmöglichkeiten. Doch auch wenn es hier künftig KI-unterstützte Hilfe gibt, hängt die Bearbeitungsgeschwindigkeit und -menge weiterhin vom Faktor Mensch ab, wenn wir saubere und verlässliche Daten anstreben. Und diese Erwartung sollten wir niemals aufgeben.

Die Bearbeitungsgeschwindigkeit wird zudem von den eingesetzten Hilfsmitteln beeinflusst. Obwohl Erweiterungen des digitalen Werkzeugkastens um neue Tools eigentlich die Arbeitseffizienz steigern sollen, erhöhen sie den Arbeitsaufwand vorübergehend deutlich. Zuweilen scheint es wie ein Fass ohne Boden. Jede neue technische Unterstützung erfordert selbst wieder, erlernt, adaptiert und angewendet zu werden. Wenn immer neue Tools

29 Vgl. Michael Piotrowski: Digital Humanities: An Explication, in: INF-DH-2018, hg. von Manuel Burghardt und C. Müller-Birn, Bonn 2018, S. 2 DOI: 10.18420/infdh2018-07.

30 Vgl. Thaller, Grundlagen (Anm. 1), S. 4. Vgl. auch das Vorwort zur ersten Ausgabe der ZfdG: Constanze Baum und Thomas Stäcker: Methoden – Theorien – Projekte, in: Grenzen und Möglichkeiten der Digital Humanities, hg. von dens., 2015 (Sonderband der Zeitschrift für digitale Geisteswissenschaften 1) DOI: 10.17175/sb001_023.

31 Vgl. CATMA-Computer Assisted Text Markup and Analysis (https://catma.de/) sowie forText. Literatur digital erforschen: CATMA (https://fortext.net/tools/tools/catma) (Zugriff: 4. Mai 2023).

32 Vgl. Heurist (https://heuristnetwork.org/, Zugriff: 4. Mai 2023).

hinzukommen, oder aber eigene neue Tools entwickelt werden, kann es dazu kommen, dass die eigentliche Datenaufbereitung immer weiter in den Hintergrund rückt. Gleiches gilt, wenn bestehende Infrastruktur angepasst werden muss. So ist jetzt schon absehbar, dass die notwendige Überführung der Texttypenklassifikation in ein Linked Open Vocabulary (LOV) Mehrarbeit bedeutet, ohne dass dadurch die Textkorpora quantitativ vorangebracht werden. Dem gegenüber stehen Mehrwerte für die Forschung in einer erweiterten Auffindbarkeit relevanter Quellen und für die Digital Humanities durch Verlinkung, Erreichbarkeit und der Erzeugung maschinenverständlicher Sinnhaftigkeit unserer Daten. Aber auch dies erfordert zunächst, sich in die Funktionsweise von LOVs und Ontologien einzuarbeiten, das Datenmodell für die Klassifikation anzupassen und mit bestehenden Schemata abzugleichen.[33]

Weiterhin kommt es vor, dass Tools nicht weiter gepflegt werden, wie es bei dem bis 2016 entwickelten *Diacollo* zu beobachten ist. Dieses Analysewerkzeug ist für die Geisteswissenschaften von hohem Nutzen, wie einige Analysen zu Zeitungskorpora demonstriert haben.[34] Wer es nutzen möchte, findet Informationen hierzu auf der offenbar auch nicht mehr aktiv betreuten Seite von Clarin-D[35] sowie als Implementierung im Digitalen Wörterbuch der deutschen Sprache (DWDS), in der Textsammlung GEI-Digital-2020 des Leibniz-Institut für Bildungsmedien, dem Georg-Eckert-Institut sowie im LTA.[36] Dieses Grundlagentool verfügbar zu halten, geht nur über weitere Projektentwicklung, was aber auch wiederum nur den Projektzeitraum absichert. Einen anderen Weg geht READ, ein ehemaliges Horizon-2020-Projekt, das seit 2019 als europäische Kooperative die Pflege und Weiterentwicklung des Transkriptionstools *Transkribus* kommerziell betreibt.[37] An diesem Beispiel lässt sich erkennen, dass digitale Techniken einige Jahre brauchen, bis sie

33 Vgl. Linked Open Vocabularies (https://lov.linkeddata.es/dataset/lov/, Zugriff: 4. Mai 2023).

34 Vgl. Daniel Burckhardt, Alexander Geyken u.a., Distant Reading in der Zeitgeschichte. Möglichkeiten und Grenzen einer computergestützten Historischen Semantik am Beispiel der DDR-Presse, in: Zeithistorische Forschungen 16/1, 2019, S. 177-196.

35 Die letzten Einträge scheinen aus dem Frühjahr 2021 zu stammen, vgl. https://www.clarin-d.net/de/aktuelles (Zugriff: 4. Mai 2023). Eine direkte Anfrage an den Clarin-D Helpdesk blieb bisher unbeantwortet.

36 Vgl. Digitales Wörterbuch der deutschen Sprache (https://dwds.de) sowie *Diacollo für GEI-Digital* (https://diacollo.gei.de/). (Zugriff: 4. Mai 2023).

37 Vgl. Read Coop. Unsere Geschichte (https://readcoop.eu/de/our-story/, Zugriff: 4. Mai 2023).

in den Geisteswissenschaften ankommen. READ ist 2016 gestartet und kann jetzt mehr als 100.000 registrierte Nutzer:innen seiner Dienste vorweisen.[38]

Ein weiterer Grund für Frustration ist die Bereitstellung von Daten, die nicht den FAIR-Prinzipien entspricht, das heißt wenn Daten entweder in den berühmt-berüchtigten Datensilos unerreichbar und nicht vernetzbar sind,[39] oder aber wenn Daten nur über verschlungene Pfade, also über diverse Weiterleitungen im Netz, erst zu erreichen sind.[40] Womöglich sind diese dann auch noch unzureichend dokumentiert oder – fast am schlimmsten – standen bereits einmal zur Verfügung und tun es aber wegen technischer Mängel oder Überalterung nicht mehr, was ebenfalls den FAIR-Prinzipien zuwiderläuft.[41]

Am wichtigsten sind jedoch die Zielvorstellungen. Ist das Produkt nur ein Testlauf? Soll es ein Ergebnis in sich darstellen? Soll es ein Hilfsmittel für eigene weitergehende Forschung sein? Oder soll es gar ein Forschungsinstrument für die Allgemeinheit werden? Gerade im letzten Fall gibt es hohe Erwartungen seitens der Nachnutzenden, also meist der Fachcommunity. Von digitalen Forschungsinstrumenten – Bibliografien, Wörterbücher oder Datenbanken – kann und muss man dauerhafte Erreichbarkeit, Stabilität, Bedienbarkeit und Nützlichkeit erwarten. Schließlich sind wir auf diese in der täglichen Arbeit ebenso angewiesen wie auf E-Mail-Programme und Officeanwendungen. Angebote, die diese Kriterien nicht erfüllen, werden schnell zur ›digitalen Leiche‹. Wer einmal auf einer Projektwebseite gewesen ist, auf der kein Fortschritt und auch kein Nutzen erkennbar ist, wird auch nicht zu dieser Webseite zurückkehren. Dies wird besonders dann gefährlich, wenn die Fachkolleg:innen ohnehin der Technik mit Skepsis begegnen. Leicht wird

38 Vgl. den Zähler auf der Startseite von READ. Dies zeigt auch, wie eine Beteiligung von Bürger:innen in Form von Citizen Science an der Aufarbeitung des Kulturerbes durch digitale Techniken und damit zugleich eine Öffnung der Wissenschaft erreicht werden kann.

39 Dies gilt insbesondere für kommerzielle Angebote wie das Corpus Christianorum Series Latina von Brepols.

40 Ein Beispiel ist das DFG-Erschließungsprojekt zu den über 10.000 historischen Urkunden des Ulmer Stadtarchivs (https://stadtarchiv.ulm.de/projekte/urkunden/projektbeschreibung, Zugriff: 4. Mai 2023). Die Digitalisate lassen sich weder in der Deutschen Digitalen Bibliothek noch im Archivportal so einsehen, dass man die Urkunden lesen könnte. Erst über eine Weiterleitung aus dem Archivportal zum DFG-Viewer erhält man eine lesbare Ansicht, aber keine Möglichkeit des Downloads, um die Bilder mit HTR lesen zu lassen.

41 Dies ist bei den älteren Digitalisaten des Albertus-Magnus-Instituts, die jetzt durch das Trier Center for Digital Humanities informationstechnisch aufbereitet und an die aktuellen Standards angepasst werden.

dann aus Enttäuschung Desinteresse oder gar Ablehnung. Gerade die Mittelalterforschung mit ihrer starken philologischen Prägung und ihrem Blick für das Detail und den Einzelfall konnte wortstatistischen Methoden lange Zeit nichts abgewinnen. Zu groß war die Skepsis gegenüber der textuellen Verlässlichkeit und dem Nutzen der Methode an sich. Dies hat sich in der Zwischenzeit glücklicherweise geändert. Doch eine gestiegene Akzeptanz heißt längst nicht, dass Distant-reading-Techniken mittlerweile einen festen Platz in der Mediävistik gefunden haben. Für deren selbstverständliche Nutzung mangelt es noch an entsprechenden Angeboten, die in ihrem Gebrauch ähnlich solide sind wie digitale Nachschlagewerke.

Um Negativspiralen zu vermeiden und stattdessen den Nutzen digitaler Angebote zu voller Geltung zu bringen, scheint es einerseits nötig, dass Erwartungshaltungen seitens der Geisteswissenschaften klarer formuliert werden, zumindest ein Grundverständnis für digitale Methoden vorhanden ist und mehr Geduld während der technischen Entwicklungsphase geübt wird. Andererseits müssen die Digital Humanities Wege finden, ihre Entwicklungen benutzerfreundlich, stabil und so nachnutzbar wie möglich zu machen. Auch sollte damit gerechnet werden, dass Technologien mit fünf bis zehn Jahren Verzögerung in einer anderen wissenschaftlichen Community ankommen. Diese Phasenverschiebung zwischen Entwicklung und Anwendung ist vielleicht eine der großen Hürden für die Digitalisierung der Geisteswissenschaften. Denn es versteht sich von selbst, dass gerade die Informationswissenschaften nicht so lange warten können, bis die Geisteswissenschaften eine Technik rezipiert und übernommen haben. Ob sich die Rezeptionsgeschwindigkeit erhöhen lässt, hängt wohl wiederum von Nutzen, Bedienbarkeit und Verlässlichkeit digitaler Tools ab, aber auch einer hoffentlich weiter steigenden Offenheit und Digital Literacy seitens der Geisteswissenschaften. Davon dürfen wir ausgehen, da wir mittlerweile die vierte Phase der Digitalisierung erreichen. Handelt es sich bei den ersten drei Phasen nach Andreas Kuczera um Vollbilddigitalisierung, Volltextdigitalisierung und der Vernetzung und Annotation[42] (was *mutatis mutandis* auch für andere Medien gilt), folgt nun die vierte Phase mit der Volltextanalyse vernetzter Daten. Es klingt zwar wieder nach einem Versprechen, doch ist schon abzusehen, dass wir auf der Ebene der Analysen zu einem Mehrwert an Erkenntnis kommen werden.

42 Vgl. Jürgen Warmbrunn: Tagungsbericht: Digitale Edition und Generierung von Forschungsdaten, in: H-Soz-Kult, 14. Februar 2023, http://www.hsozkult.de/conferencereport/id/fdkn-133784 (Zugriff: 4. Mai 2023).

Damit dies aber gelingen kann und Techniken Zeit haben sich zu etablieren, bedarf es großer Forschungsverbünde, die für die institutionelle Absicherung und den fortlaufenden Betrieb inklusive Redaktionen sorgen. Daher ist es so lobenswert, dass sich im MWW-Forschungsverbund Institutionen zusammengeschlossen haben, von denen wir erwarten können, dass sie ihre Angebote langfristig zur Verfügung stellen werden. Denn nur so können wir die Phasenverschiebungen und unterschiedlichen Geschwindigkeiten der beteiligten Wissenschaften auffangen und den vollen Nutzen der digitalen Aufbauarbeit der letzten 20 Jahre ernten. Wenn nun die beteiligten Wissenschaften mit der Förderinstitutionen Wege und Instrumente jenseits kurzfristiger Projektfinanzierung finden, steht einer gelungenen Digitalisierung der Geisteswissenschaften nichts mehr im Wege.

Vielleicht werden wir uns dann zuweilen immer noch wie Gefangene im Hamsterrad des Preprocessing fühlen. Motivierender kann es dann sein, sich nicht nach der Himmelsleiter ins Glück zu sehnen, sondern an die vielen Medien- und Wissensrevolutionen vor uns zu denken und wie viel Zeit und Aufwand es kostete, bis diese sich durchgesetzt hatten.[43] Falls dann doch noch einmal Frust aufkommt, kann der Gedanke trösten, dass wir alle wie die mittelalterlichen Wissenstradenten nichts anderes sind als Arbeiter:innen im Weinberg des Herrn.

43 Vgl. Stefan Tanaka: Old and New of Digital History, in: History and Theory 61,4, 2022, S. 3-18 für die lange Phase der Digital History 1.0, die er mit Roberto Busa beginnen lässt. Sowie Michael Giesecke: Von den Mythen der Buchkultur zu den Visionen der Informationsgesellschaft. Trendforschungen zur kulturellen Medienökologie (Suhrkamp Taschenbuch Wissenschaft 1543), Frankfurt am Main 2002, S. 270-330.