

The Glossarium Graeco-Arabicum

(Proposal for the Jahrestagung der Digital Humanities im deutschsprachigen Raum)

Applicants:

1) Yury Arzhanov

Contact: Ruhr-Universität-Bochum, Fakultät für Philologie, Sem. für Orientalistik, Universitätsstr. 150, GB 2/34, Bochum, 44780 Deutschland; e-mail: yrarzhanov@gmail.com.
Scientific interests: Ancient translations of the Greek philosophical and scientific works into Syriac and Arabic.

2) Torsten Roeder

Contact: Berlin-Brandenburgische Akademie der Wissenschaften, TELOTA, Jägerstraße 22/23, 10117 Berlin; e-mail: roeder@bbaw.de.
Scientific interests: Digital Humanities, Musicology, Italian Language and Literature.

The Project

From the eighth to the tenth century A. D., Greek scientific and philosophical works were translated to large extent into Arabic. This activity resulted in the incorporation and reorganization of the classical heritage in the new civilization which, using Arabic, spread with Islam.¹ The object of project *Glossarium Graeco-Arabicum* is to make readily available to scholars the direct information which the Graeco-Arabic translations contain for several areas of research. The *Glossarium Graeco-Arabicum* is hosted by the Ruhr-Universität Bochum, starting with DFG funds in 1994, and was continued within the ERC project “Greek into Arabic - Philosophical Concepts and Linguistic Bridges” since 2010.²

Digital Resources

The database *Glossarium Graeco-Arabicum* makes available the files of a lexical project, intended to open up the lexicon of the mediæval Arabic translations from the Greek.³ It contains images of ca. 80,000 filecards which have not yet been published in the analytical reference dictionary A Greek and Arabic Lexicon,⁴ and comprises Arabic roots from the letter *jîm* to the end of the Arabic alphabet. The database provides search facilities for Greek words, Arabic words and roots, as well as the authors and titles of the source texts. It is thus a possible basis for generating entries in Greek-Arabic dictionaries. To extend the effectiveness of the database, it is intended to link the data with the Perseus Digital Library and other online resources as well.

Technical Aspects

The coexistence of several alphabetic systems within one web application brings to light a number of phenomena and issues, as competing encoding systems, concurring writing

¹ <http://www.ruhr-uni-bochum.de/rubin/rubin-fruehjahr-2012/pdf/beitrag2.pdf> (2013-08-27)

² <http://www.greekintoarabic.eu/> (2013-08-27)

³ <http://www.ruhr-uni-bochum.de/imperia/md/content/orient/glossarium-graeco-arabicum.pdf> (2013-08-27)

⁴ A Greek and Arabic Lexicon, Leiden: Brill, 1992ff.

directions, and characters of very different and individual types. Difficulties usually do not occur in single alphabet environments, but when more than one writing systems are used parallelly in the same context, some seemingly trivial problems arise again. This concerns the representation of research contents in the database (backend), the resulting website and its input methods (researcher backend), and public access and search methods as well (researcher frontend). While Unicode seems to be the universal solution, there are still some issues that require complicated workarounds, or result in disadvantages for the user otherwise.

Disambiguierung in Suchtrefferlisten aus großen Textkorpora: Anwendungsfelder und Perspektiven

Thomas Bartz¹, Alexander Geyken³, Christian Pöltz², Achim Saupe⁴, Angelika Storrer¹
Technische Universität Dortmund, Institut für deutsche Sprache und Literatur¹ / Fakultät Informatik²
Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), Zentrum Sprache³
Zentrum für Zeithistorische Forschung Potsdam (ZZF)⁴

1. Zielsetzung und Projekthintergrund

Digitale Textkorpora bieten in vielen geisteswissenschaftlichen Arbeitsbereichen neuartige Möglichkeiten, Forschungsfragen an authentischen Sprachverwendungen zu untersuchen. Infrastrukturprojekte wie CLARIN bieten flexible Werkzeuge zur Datengewinnung und zur quantitativen Analyse an, mit denen große, linguistisch strukturierte Textkorpora ausgewertet werden können. Allerdings müssen die automatisch gewonnenen Daten oft noch manuell nachbearbeitet werden. Dies ist insbesondere der Fall, wenn nicht Wortformen, sondern sprachliche Zeichen – also Verbindungen von Form und Inhalt – quantitativ ausgewertet werden sollen, denn homonyme und polysemic Textwörter sind in aktuell verfügbaren Korpora nicht disambiguiert. Wenn die Unterscheidung von Homonymen und polysemic Lesarten für eine Forschungsfrage relevant ist, müssen die Daten bislang manuell disambiguiert werden. Der damit verbundene Aufwand ist oft erheblich; unter den zeitlichen Restriktionen eines Forschungsprojekts, einer Dissertation, einer studentischen Abschlussarbeit etc. können so bestimmte Fragestellungen gar nicht bearbeitet werden.

Im Verbundprojekt „Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining“ (KobRA, <http://www.kobra.tu-dortmund.de>) arbeiten germanistische Linguistik, Informatik, Sprachtechnologie und Sprachressourcenanbieter gemeinsam daran, den Aufwand der manuellen Nachbearbeitung zu senken und damit die Möglichkeiten der korpusbasierten Recherche und Analyse zu verbessern.¹ Dazu werden Machine-Learning- und Data-Mining-Verfahren des Informatikpartners für Aufgaben aus aktuellen Forschungsvorhaben der Linguistik angepasst und in Fallstudien erprobt. Die beteiligten Sprachtechnologie- und Sprachressourcenpartner stellen dazu unterschiedlich strukturierte große digitale Textkorpora bereit (z.B. wortartenannotierte Korpora, Baumbanken etc.) und integrieren die entwickelten Verfahren in die vorhandene Infrastruktur. Die Fallstudien beziehen sich auf drei Anwendungsfelder: Korpusbasierte Lexikographie, diachronische Sprachforschung und Varietätenlinguistik. Bei den daraus abgeleiteten Aufgabenstellungen handelt es sich um Routineaufgaben bei der Arbeit mit großen Textkorpora (Filtern, Klassifizieren, Disambiguieren, Visualisieren), die sich in verschiedenen geisteswissenschaftlichen Arbeitsbereichen in ähnlicher Form stellen.

Im Vortrag erläutern wir zunächst ausgehend von einem konkreten Anwendungsszenario aus der korpusbasierten Lexikographie Herausforderungen, die bei der Arbeit mit großen Textkorpora entstehen, und leiten daraus Anforderungen an mögliche automatische Verfahren ab. Im Anschluss stellen wir erste erprobte Verfahren, verwendete Korpora und bisher erzielte Ergebnisse vor. In einem dritten Schritt zeigen wir schließlich perspektivisch den analytischen Mehrwert der Verfahren auch für Aufgabenstellungen der historischen Forschung auf.

¹ Das Verbundprojekt wird vom Bundesministerium für Bildung und Forschung (BMBF) seit Herbst 2012 im Rahmen des Programms „eHumanities“ gefördert. Beteiligt sind die folgenden Institutionen und Projektleiter: TU Dortmund (Germanistik: Angelika Storrer, Informatik: Katharina Morik), Berlin-Brandenburgische Akademie der Wissenschaften (Alexander Geyken), Eberhard-Karls-Universität Tübingen (Erhard Hinrichs), Institut für deutsche Sprache, Mannheim (Marc Kupietz/Andreas Witt).

2. Fallstudie im Anwendungsfeld korpusbasierte Lexikographie

2.1 Anwendungsszenario

Ein wichtiges Einsatzgebiet für digitale Textkorpora ist seit langem die Sprachlexikographie (vgl. Engelberg/Lemnitzer 2009). In einem digitalen Referenzkorpus wie dem DWDS-Kernkorpus (vgl. Geyken 2007), das im Hinblick auf die Verteilung der enthaltenen Textbestände auf die Textsortenbereiche Belletristik, Gebrauchsliteratur, Wissenschaft und journalistische Prosa sowie auf die Dekaden des 20. Jahrhunderts ausgewogen ist, können Lexikographen zu einem Suchwort automatisch Daten zur Frequenzentwicklung über das 20. Jahrhundert hinweg gewinnen und die Gebräuchlichkeit des Wortes in verschiedenen Textsortenbereichen vergleichen. Wenn man allerdings Aussagen zur Textsortenspezifität und zur Bedeutungsentwicklung einer speziellen Wortbedeutung treffen möchte, müssen die vom System ausgegebenen Belege bei polysemen oder homonymen Lexemen manuell disambiguiert werden. Wenn sich die Anzahl der Treffer zu einem Suchwort in überschaubaren Grenzen hält, wie im Falle des in Storrer (2011) diskutierten Beispielworts *Ampel*, ist eine solche Disambiguierung noch mit vertretbarem Zeitaufwand möglich. Bei dem in 2.3. untersuchten Beispielwort *Leiter* resultiert die Suche im DWDS-Kernkorpus bereits in einer Liste mit 6895 Belegen, die in nicht vorhersehbaren Anteilen Belege für die Homonyme *der Leiter* und *die Leiter* sowie für speziellere Lesarten (z.B. *Leiter* i.S. von *Energieleiter*, *Trittleiter*, *Tonleiter*) enthält. Für viele Lexeme sind die Belegzahlen noch höher; eine manuelle Disambiguierung ist in solchen Fällen zeitlich extrem aufwändig. In unserer unten beschriebenen Fallstudie suchen wir deshalb nach Verfahren zur automatischen Disambiguierung von Suchwörtern in Belegen, wie sie von Korpusrecherchesystemen ausgegeben werden. Die Verfahren sollen das Arbeiten mit Korpora in der Lexikographie vereinfachen und verbessern. Weiterhin sollen auf ihrer Basis statistische Analyse- und Visualisierungswerzeuge für Korpusdaten (z.B. Kookkurrenzanalysen, Wortverlaufsdiagramme), die bislang noch überwiegend formbasiert arbeiten, um Komponenten zur semantischen Disambiguierung angereichert werden. Zudem sollen Verfahren entwickelt werden, die auch für Suchwörter, die als monosem gelten, ungewöhnliche und/oder neuartige Verwendungen zu Tage fördern.

2.2 Verwandte Arbeiten

Das vorgestellte Anwendungsszenario liegt in Reichweite der Forschung zur automatischen Disambiguierung von Wortbedeutungen (Word-Sense-Disambiguation, WSD), der sich zahlreiche Arbeiten widmen, die hier nicht in ihrer Breite dargestellt werden können (für einen Überblick vgl. Agirre et al. 2007; eine umfangreiche Vergleichsstudie zu aktuellen Verfahren hat Navigli 2009 veröffentlicht). Sie scheinen in unserem Fall auch nicht zielführend zu sein, weil sie i.d.R. Wörter nach vorgegebenen Lesarten disambiguieren. Dadurch wäre aber die Möglichkeit, im Korpus potenziell enthaltene unerwartete Lesarten zu entdecken, von vornherein ausgeschlossen.

Wir folgen in unserer Studie daher einem Ansatz, bei dem die Lesarten induktiv aus den Korpusdaten ermittelt werden. Für diese Word-Sense-Induction (WSI) liegen eine Reihe erfolgreicher Ansätze vor, die im Wesentlichen auf Clustering-Verfahren basieren (für einen Überblick vgl. Brody/Lapata 2009). Brody und Lapata (2009) konnten zeigen, dass sich mithilfe der Latent-Dirichlet-Allocation (LDA, vgl. Blei et al. 2003) tendenziell die besten Ergebnisse erzielen lassen. LDA basiert auf der Annahme, dass dasselbe Wort in verschiedenen Lesarten verwendet wird, wenn es in unterschiedlichen Kontexten vorkommt. Dazu werden um alle Vorkommen eines zu behandelnden Wortes Kontextfenster in einer bestimmten Größe gelegt und mithilfe von Wort- und Kookkurrenzstatistiken Verteilungen von Kontextwörtern, sogenannte „Topics“, ermittelt, die als Lesarten aufgefasst werden können. Für jedes einzelne Kontextfenster lässt sich daraufhin die Wahrscheinlichkeit berechnen, mit der es einem bestimmten Topic bzw. ein Vorkommen des zu behandelnden Wortes einer bestimmten Lesart zugeordnet werden kann. Dabei wird angenommen, dass die Wahrscheinlichkeit für die Zuordnung zu den

Topics einer Dirichletverteilung folgt. Rohrdantz et al. (2011) zeigten den Nutzen des Verfahrens als Grundlage für Visualisierungen zur Bedeutungsentwicklung von Beispielwörtern aus einem Zeitungskorpus, die es erlauben, die Entstehung von Neubedeutungen und ihre Entwicklung über die Zeit zu rekonstruieren.

2.3 Erste Ergebnisse

Für die im Vortrag vorgestellte Fallstudie wird LDA an Sprachdaten aus dem DWDS-Kernkorpus des 20. Jahrhunderts (s. 1.) erprobt. Ergänzend zu Rohrdantz et al. (2011) können so weitere Erkenntnisse über den Nutzen des Verfahrens auch für deutsche Sprachdaten gewonnen werden, die zudem aus unterschiedlichen Textsortenbereichen stammen. Bislang wurde das Verfahren am Beispiel des Wortes *Leiter* evaluiert (6895 Belege aus dem DWDS-Kernkorpus, 2000 manuell disambiguerte Belege für die Evaluation; Verteilung der Lesarten s. Tabelle 1). Die Topics wurden zunächst ausschließlich auf

Lesart	Vorkommen	
	absolut	relativ
<i>Boss/Führungs person</i>	1665	0,833
<i>Trittleiter</i>	296	0,148
<i>Energieleiter</i>	29	0,015
<i>Tonleiter</i>	10	0,005
	2000	

Tabelle 1: Manuell ermittelte Lesarten

Basis der Kontextwörter (Bag-of-Words) in einem noch verhältnismäßig großen Fenster im Umfang des jeweils ganzen das Wort *Leiter* enthaltenden Satzes ermittelt. Zur Evaluation des Verfahrens wurde die Reinheit der Topic-Cluster im Hinblick auf die manuell bestimmten Lesarten gemessen, als Maß diente NMI (Normalized Mutual Information, vgl. Manning et al. 2008: 357 f.).

Wenngleich der angewandte Ansatz ($NMI = 0,20$) einem einfachen k-Means-Clustering (vgl. Lloyd 1957/1982; $NMI = 0,16$) bereits überlegen zu sein scheint, erfordert das beschriebene Anwendungsszenario (s. 2.1) eine weitere Verfeinerung des Verfahrens. Tabelle 2 veranschaulicht die Ergebnisse: Zum einen lassen sich die ermittelten Topics nur den zwei häufigsten manuell bestimmten Lesarten zuordnen (*Boss/Führungs person i.S.v. politischer Leiter, DDR/Drittes Reich*: Topic 1/2; *Leiter einer Bildungsinstitution*: Topic 3; *musikalischer Leiter*: Topic 4; *Trittleiter*: Topic 5) – Belege für *Leiter i.S.v. Energie-* bzw. *Tonleiter* sind allerdings auch selten (s. Tabelle 1). Zum anderen sind die ermittelten charakteristischen Kontextwörter noch verhältnismäßig allgemein und ihre Salienz zu gering (z.B. im Vergleich zu syntaktischen Kookkurrenzstatistiken, vgl. Didakowski/Geyken 2013). Zur Verbesserung des Verfahrens werden deshalb gegenwärtig unterschiedliche Kontextfenster getestet sowie weitere Featureklassen (POS, Dependenz) integriert. Eine ausführlichere Beschreibung und Auswertung der getesteten Ansätze würden wir sehr gerne im Paper bzw. im Vortrag präsentieren.

Topic 1	0,2329	Topic 2	0,2181	Topic 3	0,1821	Topic 4	0,1699	Topic 5	0,1969
DDR	0,0030	politisch	0,0040	Berlin	0,0027	Musik	0,0012	hinauf	0,0016
Abteilung	0,0027	Partei	0,0025	Prof.	0,0019	München	0,0011	Mann	0,0014
Regierung	0,0023	Korps	0,0019	Dr.	0,0012	New_York	0,0011	oben	0,0014
Minister	0,0016	Führer	0,0019	Hochschule	0,0009	Dirigent	0,0010	gehen	0,0012
ZK	0,0013	Arbeit	0,0017	Institut	0,0008	Oper	0,0008	Sprosse	0,0009
SED	0,0010	NSDAP	0,0010	Lehrer	0,0008	Komponist	0,0007	Wand	0,0008

Tabelle 2: Automatisch induzierte Topics und salienteste Kontextwörter (Auszug aus Top 50), jeweils mit Auftrittswahrscheinlichkeiten

3. Anwendungsmöglichkeiten des Verfahrens in anderen geisteswissenschaftlichen Disziplinen und Anwendungsbereichen: Das Beispiel „Historische Semantik des 20. Jahrhunderts“

Die Induktion von Lesarten auf Grundlage von Kontextwörtern stellt auch eine interessante Anwendungsmöglichkeit für die Geschichtswissenschaften dar. Am ZZF, mit dem die BBAW im Kontext des CLARIN-Projekts zusammenarbeitet, ist ein Projekt zur Historischen Semantik des 20. Jahrhunderts angesiedelt (vgl. Kollmeier/Hoffmann 2010, 2012; Kollmeier/Saupe im Erscheinen), welches sich mit dem Wandel von Begriffen und Topoi sowie den mit ihnen verbundenen Diskursen beschäftigt. Die quantitative Aufschlüsselung polysemer Begriffe ist auch hier ein bedeutsamer Gewinn: sie ermöglicht einen schnelleren Zugriff auf verschiedene Bedeutungen und Verwendungsweisen von historisch relevanten Begriffen und erleichtert damit die qualitative Auswertung von Textkorpora.

Im Rahmen der Kooperation zwischen der BBAW und dem ZZF soll das oben beschriebene Verfahren (s. 2.3) auf weitere Korpora angewandt werden. Insbesondere soll auf der Basis des digitalisierten Zeitungskorpus des DDR-Presseportals (Neues Deutschland, Berliner Zeitung, Neue Zeit, <http://zefys.staatsbibliothek-berlin.de/ddr-presse/>) die sprachliche Ambiguität von zentralen Begriffen untersucht werden. Das Beispiel des Begriffs *Einheit* verdeutlicht dies. Dieser stand in der DDR in verschwindendem Maße für die deutsche Einheit, seit 1946 dagegen vorrangig für die Einheit von SPD und KPD bzw. seit 1971 für die Einheit von Wirtschafts- und Sozialpolitik. Daneben tauchte der Begriff immer auch als Maßeinheit in der Produktionsberichterstattung auf. Das Disambiguierungsverfahren könnte dabei helfen, den historisch-semantischen Bedeutungswandel des Begriffs *Einheit* in der DDR empirisch auf quantitative Weise zu analysieren.

Literatur

- Agirre, E. / Márquez, L. / Wicentowski, R. (Hg.) (2007): Proceedings of the SemEval-2007. Prague, Czech Republic.
- Blei, D. M. / Ng, A. Y. / Jordan, M. I. (2003): Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Brody, S. / Lapata, M. (2009): Bayesian word sense induction. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09. Stroudsburg, PA, USA, 103–111.
- Didakowski, J. / Geyken, A. 2013: From DWDS corpora to a German Word Profile – methodological problems and solutions. In: Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network „Internet Lexicography“. Mannheim: Institut für Deutsche Sprache (OPAL - Online publizierte Arbeiten zur Linguistik X/2012), 43–52.
- Engelberg, S. / Lemnitzer, L. (2009): Einführung in die Lexikographie und Wörterbuchbenutzung. Tübingen: Stauffenburg.
- Geyken, A. (2007): The DWDS corpus: a reference corpus for the German language of the 20th century. In: Fellbaum, C. (Hg.): Idioms and collocations. Corpus-based linguistic and lexicographic studies. London: Continuum, 23–40.
- Kollmeier, K. / Saupe, A. (im Erscheinen): Ausgangspunkte einer Historischen Semantik des Politischen für das 20. Jahrhundert. In: Kämper / Warnke (Hg.): Diskurs interdisziplinär. Zugänge, Gegenstände, Perspektiven (= Diskursmuster – Discourse Patterns, hg. von Beatrix Busse/Ingo Warnke). Berlin: Akademie-Verlag.

- Kollmeier, K. / Hoffmann, S.-L. (Hg.) (2012): Roundtable Discussion: Geschichtliche Grundbegriffe Reloaded? Writing the Conceptual History of the Twentieth Century. In: Contributions to the History of Concepts 7 (2), 78–128.
- Kollmeier, K. / Hoffmann, S.-L. (Hg.) (2010): Zeitgeschichte der Begriffe? Perspektiven einer Historischen Semantik des 20. Jahrhunderts. Debatte, in: Zeithistorische Forschungen / Studies in Contemporary History 7 (1), 75–114. Online: <http://www.zeithistorische-forschungen.de/16126041-Kollmeier-Hoffmann-1-2010>.
- Manning, C. D. / Raghavan, P. / Schütze, H. (2008): Introduction to Information Retrieval. Cambridge: Cambridge University Press.
- McEnery, T. / Xiao, R. / Tono, Y. (2006): Corpus-Based Language Studies. An Advanced Resource Book (Routledge Applied Linguistics). London, New York: Routledge.
- Navigli, R. (2009): Word sense disambiguation: A survey. ACM Computing Surveys (CSUR), 41 (2), 1–69.
- Lloyd, S. P. (1957/1982): Least squares quantization in PCM. In: IEEE Transactions on Information Theory, 28, 129–137.
- Lüdeling, A. / Kytö, M. (2008/9) (Hg.): Corpus Linguistics. An International Handbook. 2 Bände. Berlin, New York: de Gruyter.
- Rohrdantz, C. et al. (2011): Towards Tracking Semantic Change by Visual Analytics. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon, 305–310.
- Storrer, A. (2011): Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In: Knapp et al. (Hg.): Angewandte Linguistik. Ein Lehrbuch. Tübingen: Francke, 216–239.

Kostümsprache als Mustersprache: Vom analytischen Wert Formaler Sprachen und Muster in den Filmwissenschaften

Johanna Barzen, Frank Leymann
Institut für Architektur von Anwendungssystemen (IAAS)
Universität Stuttgart
[Barzen | Leymann]@iaas.uni-stuttgart.de

Kleidungssprachen als Formale Sprachen

In den Medienwissenschaften ist die Frage, wie Kostümsprache im Film greifbar und verstehtbar gemacht werden kann, ein Problem: Eine präzise Definition des Begriffs erweist sich als schwierig. Hier kann das Konzept einer Formalen Sprache aus der Informatik genutzt werden, um eine solche präzise Definition zu geben.

Dazu betrachten wir zunächst die Bestandteile von Kleidung (wie Hose, Hemd,...) eines Genres X (wie Western, Science Fiction,...) als ein *Alphabet* Σ_X . Nicht jede mögliche Kombination solcher Kleidungsbestandteile (d.h. ein Wort über dem Alphabet) ist eine Kleidung, die in dem Genre auftritt. Somit muss die Menge aller möglichen Wörter Σ_X^* eingeschränkt werden, um die Kombinationen von Kleidungsbestandteilen, die in dem Genre auftreten, zu filtern. In unserem Ansatz geschieht dies über *Produktionsregeln*, die angeben, wie „sinnvolle“ Kleidung aus Kleidungsbestandteilen zusammengesetzt werden kann; ein Anzug etwa besteht aus einer Hose, einer Weste und einem Sakko, in Zeichen: „Anzug → Hose Weste Sakko“. Das Alphabet und Namen von zusammengesetzter Kleidung (etwa „Anzug“) ist das *Vokabular* V_X einer Grammatik, und die von dieser Grammatik erzeugte Sprache ist die Kleidungssprache des Genres:

Definition: Eine *Grammatik* der Kleidung des Genres X ist ein Tupel $G_X = (V_X, \Sigma_X, P_X, S_X)$ mit:

- Σ_X ist ein Alphabet,
- V_X ist ein Vokabular (mit $\Sigma_X \subseteq V_X$),
- P_X ist die Menge der Produktionsregeln über V_X ,
- S_X ist das *Startsymbol* (mit $S_X \in V_X \setminus \Sigma_X$).

Eine *Produktionsregel* ist ein Paar von Wörtern über dem Vokabular V_X für das gilt:

$$(a, b) \in P_X : \Leftrightarrow a \in V_X^* \setminus \Sigma_X^* \text{ und } b \in V_X^*.$$

Statt $(a, b) \in P_X$ wird auch geschrieben: $a \rightarrow b$. Die Menge aller Wörter, die aus dem Startelement und der iterativen Anwendung von Produktionsregeln der Grammatik G_X erzeugt werden können, nennen wir die *Kleidungssprache* $L(X)$ des Genres X. □

Die Bestandteile von Kleidung eines Genres und auch die Produktionsregeln für sinnvolle, d.h. in dem Genre auftretende Kleidung wird durch genaue Analyse eines repräsentativen Filmkorpus des Genres abgeleitet. Hierfür haben wir ein System entwickelt (s.u.), in welchem sowohl die Kleidungsbestandteile als auch deren auftretenden Kombinationen, also die Kleidung, erfasst werden können.

Kostümsprachen als Mustersprachen

Aber Kleidung ist noch kein Kostüm: ein *Kostüm* ist Kleidung mit filmisch intendierter Wirkung. Die Festlegung der Wirkung von Kleidung wird in unserem Ansatz abstrakt durch eine *Wirkungsfunktion* $w:L(X) \rightarrow \{\text{wahr, falsch}\}$ repräsentiert. Die Realisierung dieser Funktion in der Praxis geschieht zum Beispiel dadurch, dass ein Schwellwert für die Häufigkeit des Auftretens einer Kleidung im Filmkorpus festgelegt wird und Kleidung, die diesen Schwellwert überschreitet, als Kostüm ausgezeichnet wird; oder ein Experte beurteilt die Wirkung. Damit können wir nun formal definieren:

Definition: Die Menge $\mathcal{K}_X = \{k \in L(X) \mid w(k) = \text{wahr}\}$ heißt *Kostümsprache* des Genres X. \square

Ein Kostüm kann als „bewährte Lösung“ eines wiederkehrenden Wirkungsproblems aufgefasst werden, ist somit ein *Muster* im Sinne der (Software-) Architekturen. Verweisen Muster aufeinander, um die Lösung eines Problems durch Verfeinerung oder Komposition zu beschreiben, spricht man in der Informatik von einer *Mustersprache*. Auch Kostüme verweisen mit verschiedenen Bedeutungen aufeinander, etwa um ihr gemeinsames Erscheinen zu beschreiben oder um auf die Kombinierbarkeit von Kostümen hinzuweisen. Damit ist eine Kostümsprache als eine Mustersprache darstellbar, und Methoden und Techniken der Informatik zum Umgang mit Mustern lassen sich nun auf Kostüme übertragen.

Informationssystem für Kleidungssprachen und Kostümsprachen

Unser System zur Erfassung von Kleidung basiert auf von uns erstellten umfangreichen Taxonomien von Kleidungsbestandteilen und deren Eigenschaften. Diese Taxonomien sind notwendig, um die Detailinformationen über Kleidung in Filmen, die durch mehrere Beobachter eingepflegt werden, vergleichbar zu machen: So gibt eine Menge von Taxonomien etwa vor, welche Art Kleidungsbestandteile möglich sind, eine weitere Taxonomie welche Farbnuancen erlaubt sind, wieder eine andere Taxonomie welche Materialeigenschaften beobachtbar sind, usw.; diese Taxonomien sind (kontrolliert) erweiterbar, um sich den Gegebenheiten eines Filmkorpus anzupassen. Im wesentlichen werden Kleidungsbestandteile und deren Eigenschaften erfasst, sowie beobachtete Kombinationen (d.h. Produktionsregeln) von Kleidungsbestandteile in Kleidung des Genres. Am Ende der Erfassung des Filmkorpus ist dessen Grammatik erstellt.

Nachdem der Filmkorpus eines Genres erfasst wurde, wird das System die Analyse der erfassten Kleidung unterstützen, um die Kostüme zu identifizieren: im Wesentlichen werden hier Varianten der Wirkungsfunktion unterstützt. So können Häufigkeitsanalysen auf Kostüme hinweisen, Experten können Kleidung begutachten und deren Wirkung bestätigen usw. Am Ende der Analyse ist die Kostümsprache des Genres erstellt.

Ausblick

Ein Beirat aus Experten wird uns bei der Bewertung unserer Forschungsergebnisse unterstützen und weitere Arbeiten steuern. So hilft etwa eine etablierte Kostümbildnerin bei den relevanten Taxonomien und Anforderungen an die Mächtigkeit der Anfragen an unser System. Ein bekannter Autor von Mustersprachen achtet auf die mögliche Nutzung unseres Systems für weitere Muster-Domänen.

Mag. Dr. Hanno Biber,
Österreichische Akademie der Wissenschaften,
ICLTT - Institut für Corpuslinguistik und Texttechnologie,
1010 Wien, Sonnenfelsgasse 19/8

Hanno Biber

AAC-FACKEL. Das Beispiel einer digitalen Musteredition.

Die AAC-FACKEL wurde unter Anwendung corpuslinguistischer und texttechnologischer Methodologien im Rahmen des an der Österreichischen Akademie der Wissenschaften gestarteten Forschungsprogrammes „AAC-Austrian Academy Corpus“ als digitale Musteredition eines literaturgeschichtlich überaus bedeutenden Textes konzipiert und online gestellt.¹ Das AAC ist ein Textcorpus zur deutschen Sprache zwischen 1848 und 1989, mit dem philologische Grundlagenforschung im noch relativ jungen Paradigma der computergestützten Textwissenschaften geleistet werden kann. Im Folgenden sollen als exemplarischer Anwendungsfall die aus der Corpusforschung resultierende digitale Musteredition, ihr Zustandekommen und die dafür notwendigen Bedingungen einer sich mit der Sprache und mit Fragen des Sprachgebrauchs auf empirischer Textbasis befassenden Forschungsrichtung, beschrieben werden.

Seit der Veröffentlichung der AAC-FACKEL, der digitalen Version der von Karl Kraus vom 1. April 1899 bis Februar 1936 in Wien herausgegeben satirischen Zeitschrift „Die Fackel“, haben sich bis Jahresende 2013 mehr als 25.000 Benutzer auf der Website registriert, wo die bereitgestellten Daten sowohl von den von spezifischen Text- und Sprachinteressen geleiteten Wissenschaftlerinnen und Wissenschaftlern erforscht, als auch von allgemein an der Sprache und Literatur interessierten Leserinnen und Lesern aus aller Welt vielfältig genutzt werden. Das nach den für diese Edition entwickelten Prinzipien zur Funktionalität digitaler Textressourcen und von erforderlichen Überlegungen zum grafischen Design bestimmte Interface der AAC-FACKEL ermöglicht den Benutzern, die digital aufbereiteten Texte

¹ AAC - Austrian Academy Corpus: AAC-FACKEL. Online Version: »Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936« AAC Digital Edition No. 1 (Hg. Hanno Biber, Evelyn Breiteneder, Heinrich Kabas, Karlheinz Mörth), <http://www.aac.ac.at/fackel>

sowohl lesen, als auch in komplexer Weise ihre Formen untersuchen und analysieren, sowie einfach in ihnen nach sprachlichen Einheiten und deren Eigenschaften suchen zu können.

Das Werk von Karl Kraus ist als bedeutender Beitrag der deutschsprachigen Literatur zur Weltliteratur zu betrachten und seine satirischen und polemischen Texte sind von thematischer Vielfalt, sprachlicher Komplexität und historischer Relevanz, weshalb ihre Überlieferung im digitalen Medium als unerlässlich erachtet werden kann. In der digitalen Edition der AAC-FACKEL wird die einzigartige sprachliche, literarische und satirische Qualität der Texte unter Nutzung texttechnologischer Instrumente durch verschiedene Suchmöglichkeiten und Register erschließbar gemacht. Neben der Volltextsuche und den Wortformen-Registern mit ungefähr 6 Millionen Wortformen bietet die AAC-FACKEL ein vollständiges, erstmals publiziertes Inhaltsverzeichnis sämtlicher Texte der Zeitschrift, das unter Nutzung informationstechnologischer Verfahren für die digitale Edition neu erstellt wurde. Dabei wurden sowohl die Angaben von Karl Kraus in den Überschriften der einzelnen Beiträge, bzw. von den Textanfängen und den Inhaltsangaben der Hefte, als auch jene Inhaltsverzeichnisse berücksichtigt, die vom Herausgeber nachträglich für die Quartalsbände erstellt wurden. Die vollständige Bild-Beigabe aller 22.586 Textseiten als Faksimiles ermöglicht den Nutzern die quelleneditorisch korrekte Zitierung der Texte in den 415 Heften bzw. 922 Nummern der 37 Jahrgänge der Zeitschrift. Es ist geplant, neue Funktionen wie etwa ein auf einer im AAC erstellten und bearbeiteten Namendatenbank der „Fackel“ beruhendes Personennamenregister sowie ein Verzeichnis der Varianten der Hefte der Zeitschrift in einer neuen Version der AAC-FACKEL zu implementieren.

Für die im AAC konzipierten digitalen Editionen wurde im Rahmen der texttechnologischen Forschungen ein spezifisches von Anne Burdick gestaltetes Navigationsmodul für Zeitschriften und andere Textformen entworfen, mit dessen Hilfe nicht nur von Seite zu Seite, von Heft zu Heft oder von Jahrgang zu Jahrgang navigiert werden kann, sondern auch zu im jeweiligen Zusammenhang relevanten Textpassagen. Die besondere grafische Umsetzung im Webinterface, im Bereich des Inhaltsverzeichnisses, kompensiert in diesen Fällen das Fehlen der ertastbaren physischen Objekteigenschaften und visuellen Informationen, die sonst nur durch die Wahrnehmung der Präsenz der gedruckten Zeitschrift gegeben sind. Die digitale Edition bietet auf diese Weise eine optisch wahrzunehmende Repräsentation der sonst in den von Papierqualität, -volumen, Druck- und Bindungstechnik bestimmten Eigenschaften eines Druckwerkes. In der AAC-FACKEL ist ein linguistisches Suchmodul eingerichtet den an den Texten in besonderer Weise interessierten Lesern ermöglicht, corpusbasierte Abfragen

vorzunehmen und die Basisfunktionalität von mit linguistischen Tags versehenen Ressourcen zu nutzen und so die mit Part-of-speech- und Lemma-Informationen versehenen Text zu untersuchen.

Die im AAC erstellten Prototypen geben Antwort auf die Frage, wie komplex organisierte Texte und historische literarische Zeitschriften mit einem Inventar von literarischen Formen und sprachlichen Eigenschaften in einer adäquaten und funktionellen Form im digitalen Medium so wiedergegeben werden, dass sie unter Nutzung der texttechnologischen Möglichkeiten vom wissenschaftlichen Nutzer wie auch vom interessierten Laien im neuen Medium, in das die Texte gleichsam übersetzt werden müssen, neu gelesen, interpretiert und analysiert werden können. Die zentrale Forschungsperspektive neben der Fragestellung nach den Steuerungsvorgängen und der damit verbundenen Reinterpretation und adäquaten Präsentation der Ausgangsmaterialien derartiger digitaler Editionen (scan-images, OCR-text, xml-tags, linguistic tagging, structural tagging, database-organisation etc.) liegt in der optimalen Nutzung der durch die informationstechnologische Aufbereitung der Texte gegebenen Such- und Indizierungsverfahren sowie den verschiedenen, dadurch eröffneten Zugangsmöglichkeiten zum Text sowie zu einzelnen Elementen der sprachlich, textlich und durch beschriebene Textstrukturen der Publikationsobjekte organisierten Bestände. Eine dritte zentrale corpusrelevante Forschungsperspektive liegt in der methodisch-theoretischen Reflexion über die sich durch die Kombination von editionsphilologischen Fragestellungen mit Fragen der Informationstechnologie, der Text-Technologie und der Corpus-Forschung sowie dem Interface Design sich ergebenden Konsequenzen und Forschungsansätze.

Vernetzte Korrespondenzen: Erforschung und Visualisierung sozialer, zeitlicher, räumlicher und thematischer Netze in Briefkorpora

Abstract zur Präsentation eines Posters im Themenbereich „Geisteswissenschaften und Informatik“

„In meiner hiesigen Vereinsamung lebe ich mehr als je mit den Briefen und Sendungen aus der Ferne + ich bin mit meinem Eigentlichen weniger hier als anderswo“¹, schreibt Ida Herz, die „Archivarin des Zauberers“², aus dem Londoner Exil am 20. Januar 1937 an Thomas Mann und offenbart in diesem einen Satz, wie sehr sie Exil erfährt als Vereinzelung und Verlust, als Verlust des identitätsstiftenden Kontexts, der vertrauten gesellschaftlichen Gegenwart. Zugleich verdeutlicht sie, welche besondere Bedeutung Briefe in dieser Situation, in der die etablierten sozialen, kulturellen und künstlerischen Netzwerke und Kommunikationsstrukturen von Auflösung bedroht sind, erhalten. In den Jahren der nationalsozialistischen Herrschaft, in denen allein aus dem deutschsprachigen Raum ungefähr eine halbe Million Menschen in die verschiedensten Exilländer flohen, erfüllt der Brief mehr denn je und unter existentiellen Vorzeichen seine per se netzwerkbildenden Funktionen. Dies wird schon allein durch die unüberschaubare Zahl der im Exil entstandenen Briefe untermauert. Dennoch wurden Netzwerke und speziell Korrespondenznetzwerke des Exils von der Exilforschung bislang nur in Ansätzen angegangen.³ Wesentliche Gründe dafür scheinen die quantitative Fülle des Materials und die qualitative, thematische und gedankliche Heterogenität der Briefe zu sein, die einer Erschließung und Erforschung mit traditionellen literaturwissenschaftlichen Methoden entgegenstehen. Diese des im Rahmen der Postersession zu präsentierenden Vorhabens ist, dass die Informatik die Literaturwissenschaft und Editionsphilologie bei der Bewältigung dieser Aufgabe unterstützen kann – und dies vor allem in zwei Bereichen.

Zum einen entwickeln die Informatiker⁴ in enger Abstimmung mit den am Projekt beteiligten Geisteswissenschaftlern spezielle Werkzeuge, die die Edition und Erschließung des Korpus ausgewählter Briefe deutschsprachiger exilierter Kulturschaffender aus der Zeit von 1932–1950 erleichtern sollen. So wird der geisteswissenschaftliche Erschließungs- und Annotationsprozess (Auszeichnung von Personen, Orten, Werken etc. in XML/TEI, Kommentierung, Vernetzung über Normdaten wie GND und Geodaten) unterstützt durch Werkzeuge, die, unter Verwendung von bestehenden Diensten zur Erkennung von Eigennamen⁵, Lexika und kontextsensitiven Regeln etwa Personen, Orte oder Datierungen automatisch erkennen und den Geisteswissenschaftlern entsprechende Vorschläge zur Identifizierung unterbreiten. Diese Vorschläge werden durch interaktive, lernende, webbasierte Werkzeuge bereitgestellt, die neben der Identifizierung auch

¹Friedhelm Kröll: Die Archivarin des Zauberers. Ida Herz und Thomas Mann. Cadolzburg 2001, S. 146.

²S. den gleichnamigen Titel von Krölls Studie (wie Anm. 1).

³Vgl. etwa Burcu Dogramaci/Karin Wimmer (Ed.): Netzwerke des Exils. Künstlerische Verflechtungen, Austausch und Patronage nach 1933. Berlin 2011.

⁴Aus Gründen der Lesbarkeit wird das generische Maskulinum verwendet. Hiermit sind ausdrücklich Frauen und Männer gemeint.

⁵Z.B. Stanford Named Entity Recognizer, <http://nlp.stanford.edu/>.

Hintergrundinformationen über Personen oder historische Ereignisse liefern, denn historische Briefe können eine weite Spannbreite von Themen abdecken, die oft nicht ohne geeignetes Hintergrundwissen eingeordnet werden können. Weiterhin wird ein für das Thema ‚Exil‘ spezifischer Thesaurus erarbeitet. Durch die Verknüpfung von Textpassagen mit den Einträgen dieses Exil-thesaurus entsteht ein mächtiges inhaltliches Erschließungsinstrument. Eine spätere Suche über diesen Thesaurus etwa nach ‚Exil‘, ‚Einsamkeit‘, ‚Verlust‘ oder ‚Briefkommunikation‘ liefert damit nicht nur die rein syntaktischen Treffer einer bloßen Volltextsuche, sondern auch Briefstellen wie die eingangs zitierte, welche durch die Anreicherung mit Themen erschlossen wurden.

Zum anderen entwickelt die Informatik auf der Grundlage dieser computergestützten qualitativen philologischen Texterschließung sowie der quantitativen Auswertung der Metadaten digitale Visualisierungstechniken, die in die Bereitstellung eines netzbasierten, generischen Forschungsportals für semantisch vernetzte Briefkorpora münden, das dem Nutzer vielseitig modifizierbare Suchwerkzeuge an die Hand gibt und es ihm ermöglicht, Korrespondenznetze nicht nur in ihrer sozialen, sondern auch in der zeitlichen, räumlichen und insbesondere thematischen Dimension zu erfassen und mit den unterschiedlichsten Fragestellungen an die Texte heranzutreten. Neben den naheliegenden, eher personenbezogenen Fragestellungen, wer etwa überdurchschnittlich viele Korrespondenzpartner hatte und sich durch besonders rege briefliche Aktivität auszeichnete, während andere Briefschreiber eher separiert am Rand eines Netzes standen, sollen die vier Dimensionen auch miteinander kombinierbar sein und Fragen danach zulassen, wer sich mit wem wann über welches Thema ausgetauscht hat, wie sich Themen über die Zeit hinweg entwickelt haben oder ob bestimmten Themen zu bestimmten Zeiten und/oder an bestimmten Orten eine besondere Bedeutung bzw. Aktualität zukam. Die kaskadierende Suche ermöglicht verschiedene, vom Nutzer konfigurierbare Sichten auf die Korrespondenznetze – sowohl personen- wie auch themen- und ortsbezogene Sichten, die jeweils in Bezug auf einen gewünschten Zeitraum eingeschränkt werden können –, beginnend bei eher abstrakten Darstellungen, wie man sie aus den Bereichen Data Mining und Visual Analytics kennt, über diskrete graphbasierte Darstellungen bis hin zu interaktiven Zeitleisten, Karten und weiteren Darstellungen statistischer Daten (z. B. Graphen, Tabellen, Listen), die stets zu konkreten Briefpassagen mit einstellbarer Kontextgröße bzw. den Briefen selbst führen werden. Erwartet wird, dass die auch spielerisch modulierbare Visualisierung dabei zu Fragen anregt, die sich erst durch den neuen, unverstellten Blick auf die Daten ergeben.

Das vom BMBF im Rahmen der Ausschreibung „eHumanities“ geförderte Verbundvorhaben wird durchgeführt vom Trier Center for Digital Humanities an der Universität Trier (Leitung: Dr. Thomas Burch, Dr. Vera Hildenbrandt, Prof. Dr. Claudine Moulin), dem Deutschen Literaturarchiv Marbach (Leitung: Dr. Roland S. Kamzelak) und dem Institut für Informatik der Martin-Luther-Universität Halle-Wittenberg (Leitung: Prof. Dr. Paul Molitor, Dr. Jörg Ritter).

GeoBib - Georeferenzierte Online-Bibliographie früher Holocaust- und Lagerliteratur

Frank Binder, Annalena Schmidt, Bastian Entrup, Markus Roth, Henning Lobin

Zielsetzung

Ziel des Projekts ist es, die frühen Texte der deutsch- bzw. polnischsprachigen Holocaust- und Lagerliteratur von 1933 bis 1949 bibliographisch in einer Online-Datenbank zu erfassen. So können diese frühen Texte, die in weiten Teilen aus dem kulturellen und kollektiven Gedächtnis verdrängt wurden, für die öffentliche, wissenschaftliche und didaktische Wahrnehmung erschlossen und aufbereitet werden. Ergänzt werden die bibliographischen Einträge durch inhaltliche und biographische Annotationen, Informationen zur Werkgeschichte sowie durch Georeferenzierung (Informationen zu Orten und Plätzen anhand von Kartenmaterial).

Das zu entwickelnde Web-Portal soll dabei – neben der bibliographischen Suche – auch über geographische Karten gezielt Texte zu einer bestimmten Region zugänglich machen. Dabei sollen Abfragemöglichkeiten nach räumlichen Kriterien und Attributen beliebig kombinierbar sein.

Methoden

Die frühen Texte der Holocaustliteratur werden – verbunden mit einer tiefreichenden inhaltlichen Erschließung – in einem Online-Bibliographie-Portal repräsentiert. Eine an internationalen Annotationsstandards (TEI) ausgerichtete systematische Erfassung der bis 1949 publizierten Texte, ggf. erschienener Rezensionen, der Sekundärliteratur sowie die Anreicherung durch biographische Informationen zu den Verfassern/-innen wird dabei kombiniert mit der Georeferenzierung von Metadaten und Textinhalten (Orte, Lager, Gettos etc.). Sämtliche Daten werden in einer Online-Datenbank erfasst, die den zukünftigen Nutzern den Zugriff auf die bibliographischen Daten und deren Auswertung durch innovative kartenbasierte Visualisierungen ermöglicht. Dies bildet eine wesentliche Grundlage für daran anschließende literatur- und geschichtswissenschaftliche Forschungsfragestellungen sowie für eine didaktische Nutzung dieser Zeugnisse in der schulischen und außerschulischen Bildungsarbeit.

Genutzte Ressourcen

Die aufwändige Beschaffung und inhaltliche Erschließung der frühen Holocausttexte wird als zentraler Teil der Projektarbeiten durch ein Team von Literaturwissenschaftler/innen und Historiker/innen unter intensiver Nutzung verschiedener einschlägiger Bibliotheken, Archive sowie den Kauf antiquarischer Bücher durchgeführt.

Zur Erfassung bibliographischer, literaturwissenschaftlicher und historischer Daten wird ein auf TEI-P5 basierendes XML-Schema erstellt und eine angepasste Autorenenumgebung in Oxygen XML verwendet (s. Entrup et al. 2013b).

Historisch-biographische Informationen zu Autor/innen sowie ortsbezogene Informationen werden zeitgleich zentral in einem projektinternen Redaktionswiki (Wikimedia) zusammengetragen. Somit können sie später automatisiert ausgelesen und in die Portaldatenbank übertragen werden. Über die Verlinkung von personen-, orts- und zeitbezogenen Informationen in den TEI-Dokumenten unter

Nutzung der Wiki-Einträge werden die Zusammenhänge zwischen den erschlossenen Holocaust-Texten technisch erfassbar. Informationen zu den Autoren/innen werden darüber hinaus mit der Gemeinsamen Normdatei (GND) der Deutschen Nationalbibliothek verknüpft.

Zurückgegriffen werden kann aber auch auf die im Bibliographieportal des Herder-Instituts und anderen Abteilungen gesammelten Daten und die Forschungsbibliothek (Warmbrunn 2012), in der für die Jahre zwischen 1954 und 1998 alle landesweiten und regionalen Zeitungen Polens und weiterer Nachbarländer gesammelt wurden und wo über eine eigene Zeitungsausschnittssammlung auf zahlreiche biographische und ortsbezogene Informationen zurückgegriffen werden kann.

Für die Georeferenzierung und Bereitstellung eines geographischen Suchzugriffes wird ein geographisches Informationssystem eingesetzt. Zur Aufbereitung von Karten wird für das entstehende Online-Portal ein Map-Server benötigt, der Karten und Abfragedienste sowie GIS-Funktionalitäten zur Verfügung stellt. In Bezug auf Kartenmaterial werden vorhandene Grundlagenkarten recherchiert aber bei Bedarf auch digitales Kartenmaterial erstellt.

Entstehende Ressourcen

Die frühen Texte der Holocaust- und Lagerliteratur werden in Form einer umfangreichen Bibliographie, nicht aber in einer digitalen Volltextbibliothek erschlossen. Urheberrechtsfragen spielen hier für die Projektarbeiten eine zentrale Rolle. Die Arbeitsstelle Holocaustliteratur tritt seit geraumer Zeit dagegen auf, Opfertexte als frei verfügbar anzusehen. Für den Gesamtbestand wäre eine befriedigende Rechteklärung aufgrund der jeweils notwendigen Einzelfallprüfungen nicht möglich gewesen. Jenseits der juristischen Dimension hätte eine Volltext-Digitalisierung ohne Rechteklärung aber auch einen immensen symbolischen Schaden zur Folge: Die Rechte der Opfer würden grob missachtet. Die Ermittlung und Erschließung der Texte in einer Bibliographie, die Rückholung in das kommunikative Gedächtnis ist dagegen auch den Rechten der Opfer stark verpflichtet und will helfen, dass deren frühe Zeugnisse (wieder) sichtbar werden.

Die tiefreichende qualitative Erschließung der Quelldokumente in Form eigens erstellter Annotationsdokumente (inhaltliche Zusammenfassung, Autorbiographie, Werkgeschichte, Verschlagwortung u.a.) bilden die Datengrundlage für die weiteren informationsverarbeitenden Schritte sowie das entstehende Online-Portal.

Das entstehende Webportal soll ausgewählte relevante Metadatenstandards unterstützen und einschlägige Schnittstellen zum Harvesting von Metadaten bedienen können. Ortsbezogene Informationen aus der Erschließung der Quelldokumente werden darüber hinaus mit Grundlagenkarten verknüpft und in ein geographisches Informationssystem eingepflegt.

Parallel zum Projekt entstehen überdies Qualifikationsarbeiten in der Literatur- und Geschichtswissenschaft, in denen die frühen Textzeugnisse sowie ihre Entstehungsbedingungen auch auf Grundlage der im Projekt erhobenen Daten und gewonnenen Erkenntnisse untersucht werden. Ferner werden zwei Konferenzen mit jeweils einem literatur- und einem geschichtswissenschaftlichen Schwerpunkt durchgeführt, deren Ergebnisse publiziert werden. In Einzelfällen, die besonders aussagekräftig sind und bei denen sich die Frage des Urheberrechts zweifelsfrei klären lässt, sollen auch frühe Textzeugnisse reeditiert und somit als Volltext dem Diskurs zugänglich gemacht werden.

Referenzen

Entrup, Bastian, Maja Bärenfänger, Frank Binder and Henning Lobin(2013a): IntroducingGeoBib: An Annotatedand Geo-referenced Online Bibliographyof Early German andPolish Holocaust and Camp Literature (1933–1949). Digital Humanities 2013, University of Nebraska–Lincoln, 16-19 July 2013.
<http://dh2013.unl.edu/abstracts/ab-229.html>

Entrup, Bastian, Frank Binder and Henning Lobin(2013b):
Extendingthepossibilitiesforcollaborativeworkwith TEI/XML throughtheusageof a wiki-system. In: Proceedingsofthe 1st Workshop on Collaborative Annotations in Shared Environments: metadata, vocabulariesandtechniques in the Digital Humanities, DH CASE '13. September 10 2013, Florence, Italy. <doi:10.1145/2517978.2517988>.

Warmbrunn, Jürgen (2012). "Das Vernetzen von Menschen, Daten und Systemen – Die Forschungsbibliothek des Herder-Instituts in Marburg." In: Bernhard Mittermaier (Hrsg.) Vernetztes Wissen – Daten, Menschen, Systeme. 6. Konferenz der Zentralbibliothek, Forschungszentrum Jülich 5. - 7. November 2012 (Proceedingsband).ISBN 978-3-89336-821-1 <http://hdl.handle.net/2128/4699>

Titel des Workshops

GeoHumanities: Karten, Daten, Texte in den digitalen Geisteswissenschaften

Vorschlag eines Pre-Conference Workshop zur
1. Jahrestagung der Digital Humanities im deutschsprachigen Raum,
25.03.2014- 28.03.2014 in Passau

Beschreibung des Workshops

In den Geisteswissenschaften beschäftigen sich verschiedene Fachgebiete, etwa die Literaturwissenschaft, die Linguistik, die Geschichtswissenschaft, aber auch die Archäologie und weitere, mit Fragen, die die räumliche Dimension oder Verteilung von Artefakten oder Eigenschaften im weitesten Sinne betreffen. Der Einsatz von digitalen Karten und geographischen Informationssystemen eröffnet heutzutage vielfältige neue Möglichkeiten, diese räumlichen Dimensionen zu untersuchen, zu dokumentieren und zu kommunizieren. Voraussetzung und Herausforderung dafür ist, dass die zu untersuchenden Artefakte oder Eigenschaften ebenfalls digital repräsentierbar sind. Solchen Entwicklungen und damit verbundenen Projekten und Vorhaben soll dieser Workshop zum interdisziplinären Austausch dienen.

Daher haben wir aus verschiedenen Fachwissenschaften und ihren Berührungs punkten mit der Geographie und Informatik um Beiträge gebeten. Bei der Auswahl der Beiträge ist uns besonders wichtig, dass ein geistes-, sozial- oder kulturwissenschaftliches Forschungsinteresse zu Grunde liegt und wir ein ausgewogenes Verhältnis zwischen verschiedenen solcher Disziplinen sowie der Geoinformatik und Informatik erreichen können. Die Beiträge können Themen wie

- Erhebung, Verarbeitung, Austausch geographischer Daten im geisteswissenschaftlichen Kontext,
- Verfahren der Georeferenzierung und des Geotaggings,
- Auswertung und Visualisierung geographischer Zusammenhänge,
- Erfahrungen zum Nutzen und Einsatz geographischer Informationssysteme in den DH

umfassen, sind aber nicht auf diese beschränkt. Von besonderem Interesse sind Erfahrungsberichte zu Methoden und zum Umgang mit digitalen Ressourcen, aber auch Überlegungen zu neuen theoretischen Konzepten.

Mit der Auswahl der Vortragenden ist eine Balance zwischen den oben genannten Themen angestrebt worden. Weiterhin soll die Reihe der eingeladenen Vortragenden zeigen, dass – neben prominenten internationalen Vorhaben – auch im deutschsprachigen Raum vielfältige Forschungsaktivitäten stattfinden, die sich in einem solchen Rahmen zusammenführen lassen, und für die der angestrebte Erfahrungsaustausch hinsichtlich der eingesetzten computergestützten Methoden und digitalen Ressourcen inspirierend und hilfreich sein kann.

Über die Antragsteller

Im Rahmen des Projekts GeoBib (<http://www.geobib.info>) arbeiten wir an einer georeferenzierten Online-Bibliographie der frühen Holocaust- und Lagerliteratur. GeoBib ist einer von 24

Projektverbünden in Deutschland, die im Rahmen der eHumanities-Förderlinie des BMBF gefördert werden (<http://www.bmbf.de/press/3319.php?hilite=GeoBib>).

Prof. Dr. Henning Lobin ist geschäftsführender Direktor des Zentrum für Medien und Interaktivität an der Justus-Liebig-Universität Gießen (www.zmi.uni-giessen.de/) sowie Professor für Angewandte Sprachwissenschaft und Computerlinguistik am Institut für Germanistik der Justus-Liebig-Universität Gießen (<http://www.uni-giessen.de/cms/ascl/>) Er ist Sprecher des BMBF-eHumanities-Projekt „GeoBib“ und bereits seit 2008 an den Forschungsinfrastrukturvorhaben D-SPIN und CLARIN-D beteiligt.

Bastian Entrup, M.A. ist wissenschaftlicher Mitarbeiter im Projekt „GeoBib“ am Lehrstuhl für Angewandte Sprachwissenschaft und Computerlinguistik am Institut für Germanistik der Justus-Liebig-Universität Gießen und im Projekt für die texttechnologische Aufbereitung der gesammelten Informationen zuständig.

Ines Schiller, M.Sc ist wissenschaftliche Mitarbeiterin im Projekt „GeoBib“ am Zentrum für Medien und Interaktivität der Justus-Liebig-Universität Gießen. Als Geoinformatikerin arbeitet sie an der Entwicklung des GeoBib-Informationsportales, welches bibliographische und geographische Daten und literatur- und geschichtswissenschaftlichen Annotationen zusammenführt und über kombinierte Such- und Visualisierungsmöglichkeiten erschließt.

Dipl.-Inf. Frank Binder ist wissenschaftlicher Mitarbeiter im Projekt „GeoBib“ am Zentrum für Medien und Interaktivität der Justus-Liebig-Universität Gießen. Er ist für die Projektkoordination, die Liaison mit Forschungsinfrastrukturvorhaben, Forschungsdatenmanagement und Öffentlichkeitsarbeit zuständig.

Kerstin Bischoff (Forschungszentrum L3S Hannover), Heidemarie Hanekop (SOFI Göttingen), Kerstin Brückweh (Neuere Geschichte, Universität Trier) und Nicole Mayer-Ahuja (SOFI Göttingen und Fachbereich Sozialökonomie, Universität Hamburg)

Herausforderungen und Best Practices interdisziplinärer Kooperation in den eHumanities – Erfahrungen aus dem BMBF-Verbund „Gute Arbeit“ nach dem Boom

Die Chancen interdisziplinärer Kooperationen zwischen Informatik, Sozial- und Geisteswissenschaften sind bestechend. Der offene Austausch von Wissen und Methoden und die Bündelung komplementärer Kompetenzen können neue Forschungsansätze befähigen; bspw. können mit neuen IT-Werkzeugen ganz neue Fragestellungen angegangen werden. In der Praxis allerdings sehen sich interdisziplinäre Projektverbünde, wie sie im Rahmen der Förderung der Digital Humanities entstehen, auch neuen Herausforderungen gegenübergestellt. Neben den durchaus erwarteten Verständigungsproblemen lauern schwerwiegende, oft auch zunächst unentdeckte Probleme, die zu Ängsten und Missverständnissen und folglich zu wechselseitigen Blockaden führen können.

Der vorgeschlagene Beitrag diskutiert Chancen und Herausforderungen der interdisziplinären Kooperation am Beispiel des BMBF-Projektes „Gute Arbeit“ nach dem Boom (Re-SozIT), an dem ArbeitssoziologInnen, ZeithistorikerInnen und InformatikerInnen beteiligt sind. Zunächst wird in Kürze das Verbundprojekt und Ergebnisse aus den ersten 1,5 Jahren vorgestellt. Der Verbund betritt gemeinsam Neuland bei der Entwicklung neuer Methoden der Längsschnittanalyse von qualitativen soziologischen Studien mit neuen IT-Werkzeugen. Im zweiten Teil werden die hierbei aufgetretenen Kooperationsprobleme und Herausforderungen analysiert. Gerade in der ersten Phase eines solchen Verbundprojektes werden Weichen gestellt, die über das spätere Scheitern oder den Erfolg entscheiden. Verbreitete Risiken sind z.B. das Aneinander-vorbei-entwickeln trotz bester Vorsätze oder die Nicht-Anwendung der von der Informatik entwickelten Tools durch die Sozial- und GeisteswissenschaftlerInnen oder wechselseitige Blockaden und Ineffizienz. Im dritten Teil werden erste Erfahrungen und Best Practices zur Erkennung und Überwindung solcher Hemmnisse vorgestellt.

"Gute Arbeit" nach dem Boom (ReSozIT) – ein Pilotprojekt zur Längsschnittanalyse arbeitssoziologischer Betriebsfallstudien mit neuen e-Humanities-Werkzeugen

Das Ziel des Verbundvorhabens "Re-SozIT" besteht in der IT-basierten Erschließung und Analyse qualitativer Forschungsdaten in einer Längsschnittperspektive: über 50 seit den 1970er Jahren durchgeführte arbeits- und industriesoziologische Forschungsprojekte des Soziologischen Forschungsinstituts Göttingen (SOFI) stehen zur Verfügung. Um das qualitative Material dieser Primärprojekte für themen- und zeitübergreifende Sekundäranalysen aktueller Fragestellungen aus Arbeitssoziologie und Zeitgeschichte auszuwerten, werden neuartige IT-Werkzeuge und Analyseverfahren entwickelt. Fragestellungen sind z.B. die Subjektivierung von Arbeit, „Gute Arbeit“ als Alltagspraxis (Soziologie) oder die Wahrnehmung von Arbeit im Zeichen von Arbeitslosigkeit (Geschichte).

Im Zentrum der interdisziplinären Arbeit standen bis dato das Datenmanagement, die prototypische Entwicklung von innovativen Suchwerkzeugen sowie die Anonymisierung.

- Eine zentrale Herausforderung ist der Aufbau eines geeigneten *Datenmanagements* inklusive *Metadaten*, die die Logik von heterogenen Primärprojekten hinreichend genau abbilden, um projektubergreifende Sekundäranalysen zu ermöglichen. Denn diese erfordern weitreichende Kenntnisse über Methoden und Design aber auch soziale Rahmenbedingungen der Primärprojekte. Ein wichtiges Ergebnis ist, dass cleveres Datenmanagement den Aufwand

für die Metadatenerfassung erheblich verringern kann (bspw. durch Vererbung von Informationen). Allerdings erweist sich die Festlegung von konkreten (Meta-)Datenstrukturen als sehr kritischer Punkt zwischen den Forschern aus den unterschiedlichen Disziplinen.

- Die Sekundäranalyse qualitativer Daten kann gerade im Hinblick auf das Finden von interessanten Primärdaten von modernen IT-basierten *Such- und Erschließungswerkzeugen* profitieren. Unser Material von 350.000 Seiten digitalisiertem Material aus 6000 qualitativen Interviews ist eine manuell nicht zu bewältigende Menge. Mittels informationstechnologischer Ansätze wie Topic Modelling, Sentiment-Analyse, etc. können (semi-)automatisch Themen, Terme und Meinungen extrahiert und u.a. zur Metadatenanreicherung genutzt werden, um über die Volltextsuche hinaus eine bessere Durchsuchbarkeit zu gewährleisten.
- Um das Spannungsverhältnis zwischen der notwendigen *Anonymisierung* personenbezogener Daten und dem Interesse an aussagekräftigen Forschungsdaten zu bewältigen wurde ein Konzept erstellt, welches auf den Bausteinen Risikoklassifizierung, IT-gestützte Anonymisierungsmaßnahmen und infrastrukturelle Rahmenbedingungen mit abgestuften Zugangsrechten beruht.

Für die Sekundäranalyse der soziologischen und zeitgeschichtlichen Teilprojekte ergeben sich völlig neue Analysemöglichkeiten und für die Informatik ergeben sich neuen Möglichkeiten des Zugangs zu wertvollen Datenmaterialien zur Erprobung ihrer Algorithmen.

Herausforderungen der interdisziplinären Kooperation

Zusätzlich zu den durchaus erwarteten Kommunikations- und Verständigungsproblemen ergaben sich einige unerwartete, zunächst verdeckte Herausforderungen, die u.a. mit wechselseitiger Rollenzuschreibung zu tun hatten oder mit unterschiedlichen und unbekannten Agenden und Interessen – bei gleichzeitiger gegenseitiger Abhängigkeit bei notwendigen Arbeitsschritten. Diese führen zu Koordinierungsproblemen, Blockaden und Spannungen. Dabei bilden sich je nach Thematik unterschiedliche ‚Fronten‘.

- Typisch für *Kommunikations- und Verständigungsprobleme* sind begriffliche Differenzen, die sogar verschärft werden, wenn teils gleiche Begriffe mit unterschiedlicher Bedeutung verwendet werden. Dies betrifft u.a. die Benennung von Schaltflächen in den Tools.
- *Mangelndes Verständnis* der unterschiedlichen Arbeitsweise und Methoden, des Selbstverständnisses und der Forschungstradition der anderen Disziplin(en) erschwert die Kooperation. Es treffen hier teils gänzlich verschiedene Denkschulen aufeinander. Während die Geschichts- und qualitativen SozialwissenschaftlerInnen eher Wert auf die Entfaltung von Argumentationslinien und elaborierten Diskurs legen, werden InformatikerInnen auf klare, prüfbare, validierbare und messbare Festlegungen drängen. Es soll eindeutig spezifiziert werden, unter welchen Ausgangsbedingungen sich das System bei bestimmten Benutzerinteraktionen wie verhalten soll.
- Die Gleichzeitigkeit von methodischer und inhaltlicher Innovation (Soziologie und Zeitgeschichte) und Entwicklung neuer Tools und Werkzeuge (Informatik) führt zu *wechselseitiger zeitlicher Abhängigkeit* und nur schwer zu erfüllenden gegenseitigen Erwartungen. Strukturell problematisch ist nun, dass die Anforderungen an IT-Tools für die Sekundäranalyse daher noch teils unbekannt sind. Ein zu starkes Drängen oder „Erzwingen“ wollen einer deutlichen Strukturierung mit festhaltbaren Ergebnissen kann u.U. dann zu Übervereinfachung führen. Dabei befindet sich in unserem Projekt die Informatik in einer Sandwich-Position zwischen den Anforderungen der Herkunftswissenschaft und potentiellen Sekundärwissenschaften, die wiederum recht anders gelagerte Informationsbedürfnisse haben können.
- *Gegensätzliche Anforderungen* der Disziplinen zeigen sich auch in punkto Anonymisierung: Während die ArbeitssoziologInnen als Datengeber weitreichende Anonymisierungsmaßnahmen zusammen mit Auflagen für die Nutzung befürworten, ist für die

Geschichte die absolute Nachvollziehbarkeit der Quelle unabdingbar. Die Informatik als dritter Spieler, sieht hier die schwierige Machbarkeit von zuverlässigen IT-basierten Anonymisierungsmaßnahmen. Auch kann die Offenlegung von Forschungsdaten auf Bedenken der Datenerheber hinsichtlich möglicher Kritik stoßen. Bspw. gibt es aktuell rege zeithistorische Auseinandersetzungen, die sich um die Nutzbarkeit von sozialwissenschaftlichen Daten als Quelle oder/und als Darstellung bzw. Fakten drehen.

- Versteckte oder zumindest unbekannte Agenden und mangelndes Verständnis bzw. Wertschätzung (Stichwort Fachwissenschaftler vs. Programmierer) zwischen den unterschiedlichen Disziplinen behindern die effiziente Kooperation, solange sie nicht aufgedeckt, diskutiert und ausgehandelt werden. Hier geht es letztlich um die Hegemonie zwischen den Disziplinen: wer setzt die *Forschungsagenda* bzw. -vorgehensweise fest.
- Die Befürchtung einer Art feindlicher Übernahme durch die Informatik bei Sozial- und Geisteswissenschaftlern spiegelt sich u.a. in einer latenten *Angst vor der Automatisierung*, dem nur bedingt vorhandenen Vertrauen in die Technologien und dem Unbehagen aufgrund einer potentiellen Entmündigung durch die IT. Diese Skepsis zeigt sich insbesondere anhand der Frage, ob Strategien wie *text mining* oder *opinion mining* Such- oder Analysefunktionen beinhalten – eine Frage, die nicht zuletzt die Selbstverständnisse der beteiligten Geistes- und Sozialwissenschaften berührt: welche Rolle spielen aggregierte Daten in vorliegendem Material, können sie als „repräsentativ“ oder auch nur „exemplarisch“ gelten? Inwiefern wird die Forschungsstrategie der Primärstudien in diesen Strategien „aufgehoben“ oder „gelöscht“?
- *Divergierende Interessen* speisen sich aus dem Spannungsverhältnis zwischen den jeweils eigenen Aufgaben und Zielen der Fachwissenschaften und Kooperationsanforderungen. Zumeist sind hier Doppelanforderungen zu erfüllen, denn zeitaufwändige Arbeiten ‚nur‘ für das Projekt bzw. die andere Disziplin, für die es aber keine Meriten in der Herkunftsdisziplin gibt, können zu Störfaktoren werden. So bedeutet bspw. die Evaluierung von Software-Prototypen Aufwand für die GeisteswissenschaftlerInnen und SoziologInnen; das Nach-Programmieren von ‚trivialen‘ Features, u.U. inspiriert durch kommerzielle Programme, ist für die Informatikforschung uninteressant.

Best Practices – Erste Erfahrungen im Zueinanderfinden

Im Laufe der ersten Projekthälfte haben wir erste Lernprozesse durchlaufen. Grundsätzlich sollte die interdisziplinäre Zusammenarbeit als eigenständige Aufgabe ernstgenommen und regelmäßig evaluiert werden. Dafür ist es empfehlenswert, sich viel Zeit zum gegenseitigen Kennenlernen zu nehmen sowie eine Kultur der Offenheit und Wertschätzung zu etablieren. Das betrifft die Vorstellung und kritische Reflektion der eigenen Arbeiten, Forschungsinteressen und Projektziele, der Forschungstradition und des Selbstverständnisses der eigenen Disziplin inklusive Spielregeln innerhalb der Community sowie zentrale (aktuelle) Methoden und Arbeitsweisen. Nur auf diese Weise scheint es möglich, besser die jeweiligen Motive nachzuvollziehen, Arbeit anzuerkennen, etc.

Kompromisse und Mittelwege zu finden ist gerade in den frühen Phasen der Zusammenarbeit wichtig. Statt dass also jede Disziplin forschungsmäßig auf ihrem (hegemonialen) Maximalanspruch besteht und somit ein effektives gemeinsames Vorankommen verhindert, können erste – wenn auch u.U. technisch, sozialwissenschaftlich oder geschichtswissenschaftlich etwas weniger herausfordernde – Fragestellungen ein guter Ausgangspunkt sein, von dem aus sich dann neuer, tatsächlich interdisziplinärer Forschungsbedarf aufzeigt. In unserem Falle hat sich so eine interessante Fragestellung zum Zeitvergleich via Topic Modelling herauskristallisiert, welche konzeptionell und auch im Hinblick auf die Umsetzung weder für die ZeithistorikerInnen noch für die Informatik trivial ist.

Die Auswahl und *Anpassung der jeweils eigenen disziplinspezifischen Methoden und Vorgehensweisen* sollte soweit als möglich an den Kooperationspartnern und dem Kontext ausgerichtet

werden – mit dem Ziel gemeinsam eine neue Vorgehensweise zu erproben und ggf. neue interdisziplinäre Methoden abzuleiten. So verfolgen wir bspw. eine sehr agile Vorgehensweise der Software-Entwicklung mit vielen Iterationen, in der durchaus noch unvollständige und fehlerhafte Forschungsprototypen schnell ‚bespielbar‘ sind. Hier besteht die Gefahr, zwischen den Zyklen ‚umsonst‘ vorgearbeitet zu haben. Gerade unter dem Gesichtspunkt der teils unbekannten Anforderungen an die mit IT-Tools zu bewältigende Aufgabe bietet ein derartiges Vorgehen jedoch die Möglichkeit, nach und nach besser die Potentiale und Unzulänglichkeiten von Techniken kennenzulernen und die ‚Fantasie‘ für weitere Unterstützungsmöglichkeiten und Anforderungen anzuregen. Ausführliche Diskussionen zeigten dabei die Notwendigkeit der Anpassung von generellen IT-Methoden an die Bedarfe und Arbeitsweisen der jeweiligen Nutzungsdisziplin.

Statt Angst vor feindlicher Übernahme durch die Informatik als schwarzer Ritter auf der einen Seite bzw. die Sorge vor Giftpillen als Versuch der (vermeintlich notwendigen) Gegenwehr auf der anderen Seite sollten sich die Digitalen Geisteswissenschaften also besser als ein gemeinsames Wagnis verstehen, das für alle beteiligten Disziplinen eine Bereicherung bieten und neue Forschungsansätze hervorbringen kann.

A Flexible NLP Pipeline for Computational Narratology

Thomas Bögel

Jannik Strötgen

Christoph Mayer

Michael Gertz

Institute of Computer Science, Heidelberg University, Germany

{boegel, stroetgen, cmayer, gertz}@uni-hd.de

1 Project Overview

Temporal dependencies reveal interesting insights into the semantic discourse structure of narrative texts. The investigations of literary scientists are, as of today, mostly based on labor-intensive manual annotations. Computational Narratology, an important subtopic of the Digital Humanities, aims at facilitating annotations and supporting literary scientists with their analyses. According to Mani (2013), one aspect of Computational Narratology focuses on exploring and testing literary hypotheses through mining narrative structures from corpora. In the context of the BMBF-funded eHumanities project **heureCLÉA**, we address temporal phenomena in literary text, a genre whose temporal phenomena are different from others. For example, it is often not possible to anchor temporal expressions to real points in time, but literary texts tend to have their own time frame. Our project partners, as well as many other humanists, use CATMA, a comprehensive graphical tool for annotating data. Interfacing NLP with CATMA could drastically reduce the effort of manual annotation. The goal of heureCLÉA is to provide users with a collaborative annotation environment for tagging temporal phenomena in documents, with simple annotations (e.g., temporal expressions) being added automatically, and more complex annotations (e.g., time shifts and ellipses) being suggested. Users can correct automatic annotations, and user feedback will be used to apply machine learning techniques to improve future annotation suggestions.

In the following, we outline our flexible architecture for NLP in the domain of narrative texts, as well as promising first results for annotating the tense

of sub-sentences to demonstrate the effectiveness of our approach.

2 Architecture and Components

The heureCLÉAcorpus currently consists of more than 20 mostly German narrative texts from various authors of the 20th century. Due to the diversity of style and text characteristics, applying NLP is challenging as most systems are optimized for factual texts characterized by stable structures.

Automatically generating annotations that are related to temporal structures of texts requires information on multiple levels of the linguistic processing stack. Thus, we implemented a modular pipeline that performs annotations with increasing levels of complexity and allows for easy adaptation and exchange of different base components. We use standard off-the-shelf tools that are freely available. In order to achieve maximum flexibility and to allow for easily substitutable individual components, the pipeline is implemented as a UIMA architecture. The general pipeline architecture is shown in Fig. 1. We distinguish between general preprocessing components that are required for all subsequent narratological annotation tasks and individual machine learning modules (red background) that are tailored to one specific target annotation. We are planning to release the preprocessing stack to the research community to allow all CATMA users to perform basic NLP analyses and annotations.

2.1 Component Overview

2.1.1 CATMA Interface

The texts in our corpus are annotated by literary scientists with CATMA, a web-based collaborative

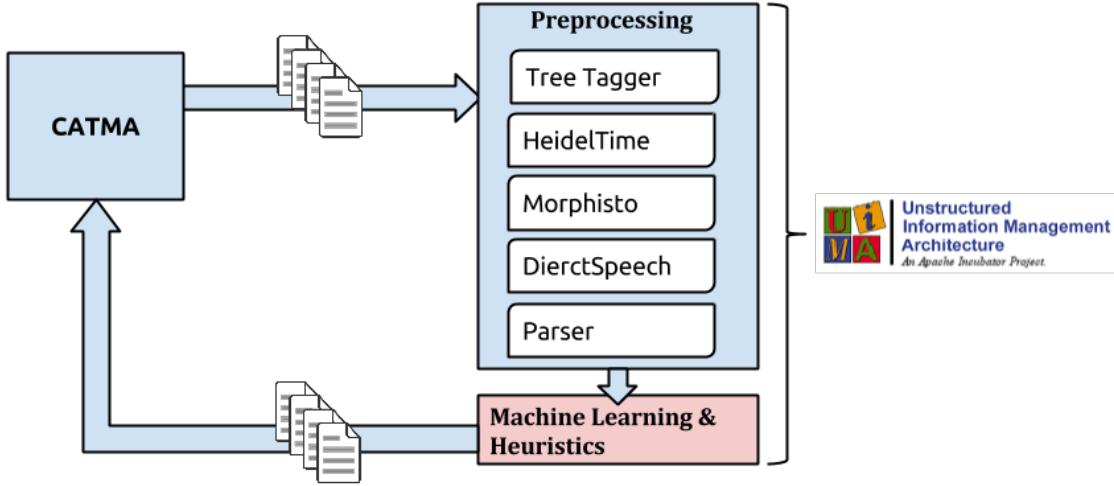


Figure 1: Overview of the NLP architecture in heureCLÉA.

annotation tool that offers humanists an easy way to create stand-off annotation and share their annotation with other scholars. In order to work with annotated data in CATMA, we implemented a component that interfaces CATMA with our UIMA pipeline. As CATMA is a popular tool in the humanities, we developed the interface as a stand-alone component that can easily be used by others to combine the strengths of CATMA as an annotation framework with the analytical and predictive power of UIMA pipelines. The interface is geared to literary scientists with no knowledge of programming. To achieve a simple configuration, the user only has to specify mappings between CATMA and UIMA types in a single XML file. By providing an easy-to-use interface, we want to lower the bar for other projects in the Humanities to employ simple yet effective NLP tools and thereby alleviate manual annotations.

2.1.2 Linguistic Processing

Our linguistic preprocessing stack consists of state-of-the-art components for German NLP. For sentence segmentation and part-of-speech tagging, we use the TreeTagger (Schmid, 1995). While the tree tagger provides a basic analysis of tokens, it does not extract morphological information; thus, we added Morphisto (Piskorski, 2009) as a separate component for morphological analysis. Two different parsers in our pipeline provide valuable informa-

tion for detecting sub-sentences and dependency relations between tokens (e.g., to extract the subject of a certain verb): the Stanford constituent parser (Raferty and Manning, 2008) and ParZu (Sennrich et al., 2009), a dependency parser trained on the TüBa-D/Z. Finally, HeidelTime (Strötgen and Gertz, 2013) extracts and normalizes temporal expressions in the text which will be used in later stages of the processing pipeline.

2.1.3 Machine Learning

After the data has been annotated by the above preprocessing components, it is passed to different modules that handle the annotation of specific narratological aspects of texts (e.g., the extraction of tense clusters or time shifts). Depending on the target annotation, we employ different machine learning approaches and heuristics.

Verb tenses are an example for such a relevant annotation because shifts in the verb tense of a sentence can, for instance, indicate narratological order phenomena (e.g., prolepsis). To extract tense clusters, we implemented and evaluated a robust heuristic component with promising results. Detailed evaluation results of the prediction performance based on a comparison to manually annotated data will be presented in the poster.

References

- Mani, I. (2013, October). Computational narratology. the living handbook of narratology. <http://www.lhn.uni-hamburg.de/article/computational-narratology>.
- Piskorski, J. (2009). Morphisto-An Open Source Morphological Analyzer for German. *Finite-state Methods and Natural Language Processing: Post-proceedings of the 7th International Workshop FSMNLP 191*.
- Rafferty, A. N. and C. D. Manning (2008). Parsing three german treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, PaGe '08, Stroudsburg, PA, USA, pp. 40–46. Association for Computational Linguistics.
- Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pp. 47–50.
- Sennrich, R., G. Schneider, M. Volk, and M. Warin (2009). A new hybrid dependency parser for german. *Proc. of the German Society for Computational Linguistics and Language Technology*, 115–124.
- Strötgen, J. and M. Gertz (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation* 47(2), 269—298.

Auf den Weg zur Visualisierung von linguistischen Mustern der Deliberation: eine Fallstudie anhand der Daten von S21

Tina Bögel Valentin Gold Annette Hautli-Janisz Christian Rohrdantz Sebastian Sulger
Miriam Butt Katharina Holzinger Daniel A. Keim

Fachbereich Sprachwissenschaft Fachbereich Politikwissenschaft Fachbereich Informatik
Universität Konstanz

1 Hinführung und Motivation

Dieser Beitrag berichtet über Fortschritte im interdisziplinären Projekt *VisArgue*¹, das sich mit der automatischen linguistischen und visuellen Analyse von politischem Diskurs befasst. Der Schwerpunkt des Projekts liegt auf der Untersuchung des Konzeptes der deliberativen Kommunikation, einer Theorie, die hauptsächlich von Habermas (1981), Dryzek (1990, 2000) Bohman (1996) und Gutmann and Thompson (1996) propagiert wird. Gemäß dieser Theorie sollten Interessenvertreter in der Kommunikation mit anderen Parteien ihre Positionen extensiv rational begründen und sich schlussendlich dem besseren Argument fügen. Die automatische Messung der deliberativen Qualität einer Gesprächssituationen erfordert die Identifikation von linguistischen Einheiten, die Aufschluss geben über Faktoren wie objektive versus subjektive Kommunikation oder dem Rückbezug auf das Allgemeinwohl bzw. demokratischen Werten als Teil der Argumentation. Konzepte wie die Haltung des Sprechers zum Gesagten und der Wahrscheinlichkeitsgehalt des Gesagten sind als rhetorische Mittel imminent wichtig, da sie konventionelle Implikaturen generieren.

Im Folgenden wird ein erster Ansatz zur Analyse von Argumentativität mittels automatischem Verfahren präsentiert. Neben einer umfassenden linguistischen Analyse der relevanten Parameter stellen wir eine computerlinguistische Implementation vor, die die Daten automatisch mit den von Seiten der Pragmatik relevanten Merkmalen annotiert. Diese Implementation kombiniert ein regelbasiertes System, das eine tiefgreifende linguistische Analyse auf die Daten anwendet, mit einem visuellen Analysesystem, das neben der visuellen Darstellung der pragmatischen Informationen eher oberflächliche Sprachmerkmale wie die Identifikation von Schlüsselworten, Modellierung der Topics der Unterhaltung, Standardberechnungen zur Länge von Sprechbeiträgen sowie die Anzahl der Wortwechsel einfließen lässt.

2 Daten

Die für die Analyse herangezogenen Daten basieren auf den Schlichtungsgesprächen zu “Stuttgart 21” (S21), das 2010 als Antwort auf den massiven öffentlichen Druck auf den Bau des Tiefbahnhofs und der damit verbundenen Stadtentwicklung in Stuttgart eingesetzt wurde. Diese als Beispiele für die Anwendung der deliberativen Kommunikation angesehenen Verfahren kommen seit den frühen 1980er Jahren in Deutschland zunehmend zum Einsatz. Die Datenbasis ist offen im Netz verfügbar² und besteht aus den transkribierten Protokollen dieser Verhandlung an neun Verhandlungstagen, jeweils mit einer Dauer von etwa sieben Stunden. Insgesamt enthalten die Transkripte etwa 265.000 Token in mehr als 1330 Redebeiträgen von 70 Personen. Die Transkripte umfassen Gespräche und Diskussionen zwischen dem Mediator, Experten, Projektbefürwortern und Projektgegnern in deutscher Sprache und wurden in ein XML-lesbares Format umgewandelt, um eine spätere automatische Verarbeitung und Annotation zu ermöglichen. Basierend auf den Informationen der online verfügbaren Transkripte wurden die Transkripte außerdem mit Sprecherinformation und Gesprächsthema annotiert.

Um eine feine Analyse des Diskurses zu erreichen, wurden alle Redebeiträge in elementare Diskuseinheiten (“elementary discourse units” — EDUs) unterteilt (Marcu 2000). Obwohl in der Literatur

¹www.visargue.uni-konstanz.de

²<http://stuttgart21.wikiwam.de/Schlichtungsprotokolle>

kein Konsensus bezüglich der genauen Eigenschaften einer EDU herrscht, wird generell angenommen, dass jede Diskuseinheit ein einzelnes Event beschreibt (z.B. Polanyi et al. 2004). Im vorliegenden Fall approximieren wir diese Annahme, indem wir alle lexikalischen Einheiten zwischen zwei Interpunktionszeichen als Diskuseinheit angesehen werden.

3 Linguistischer Hintergrund

Ein zentraler Aspekt unserer Arbeit ist die linguistisch-motivierte Operationalisierung der Kriterien die die deliberative Qualität von Kommunikation markieren, insbesondere bezüglich der Art der Realisierung und der kommunikativen Funktion von Argumenten im Diskurs und die Einstellung des Sprechers zum Gesagten. Die vorliegende Arbeit zeigt, dass dazu zwei linguistische Parameter, und zwar kausale Diskurskonnektoren und Modalpartikeln und deren Interaktion, hochrelevant sind.

Kausale Diskurskonnektoren wie *da*, *weil*, *dann*, *zumal* leiten generell die Begründung eines Sprechers ein (z.B. Prasad et al. 2008). Diese Konnektoren und deren begründungseinleitende Phrase können zwar automatisch extrahiert werden, allerdings fehlt zu einer umfassenden linguistischen Interpretation eine Antwort auf die Frage, wie die Aussage getönt ist, beziehungsweise wie forciert die Aussage kommuniziert ist und wie die Haltung des Sprechers zum Gesagten ist. Diese Faktoren werden im Deutschen insbesondere mit Hilfe von Modalpartikeln ausgedrückt (z.B. Zimmermann 2011, Karagjosova 2004). Zum Beispiel leiten *halt* und *eben* eine konventionelle Implikatur ein, die eine vom Sprecher angenommenen unabänderlichen Einschränkung durch externe Fakten ausdrückt. Diese Verwendung wird in Beispiel (1) gezeigt. Im Gegensatz dazu signalisiert *ja* wie in (2) illustriert, dass der Inhalt der Argumentation Teil des gemeinsamen Wissens der Gesprächsteilnehmer ist.

- (1) (...) weil halt in dem Bereich die meisten Autos unterwegs sind.
(...) as HALT in Art area Art most car.Pl underway be.3.Pl
'(...) because most cars are underway in this area.' (Dr. Heiner Geissler, S21, Nov. 4th 2010)
- (2) (...) da Sie ja gesagt haben, dass (...)
(...) as Pron.2.Sg.Pol JA say.Past.Part have.Inf that (...)
'(...) as you JA said that (...)' (Tanja Gönner, S21, Nov. 4th 2010)

3.1 Ambiguität

Ambiguitäten stellen ein große Herausforderung für die automatische Extraktion und Identifikation von Kausalkonnektoren und Modalpartikeln dar. Insbesondere ist dies der Fall für den Konnektor *da*, der neben dem kausalen Gebrauch auch als temporales und lokatives Pronomen fungieren kann. Allerdings kann ein Großteil dieser Ambiguitäten aufgelöst werden, indem linguistische Indikatoren wie die Position des Konnektors, seine angrenzenden Elemente und die generelle Struktur des Satzes mit einbezogen werden. Dies alles fließt in eine Disambiguierungsregel ein, wie in (3) dargestellt.

- (3) IF *da* nicht gefolgt von einem Verb AND
kein anderer Partikel oder Konnektor vor *da* AND
finales Verb ist ein Infinitiv THEN
da ist ein Kausalkonnektor.

Dieselbe Herangehensweise wird für Modalpartikeln angewendet, zum Beispiel kann *eben* neben seiner Verwendung als Modalpartikel auch noch als Fokuspartikel und temporales Adverb auftreten. Als Modalpartikel signalisiert *eben* die resignierte Zustimmung zu einer Sache aufgrund unabänderlicher Randbedingungen (Kwon 2005).

3.2 Inferenzregeln

Während die zwei oben genannten Parameter für sich genommen schon wichtig für die Interpretation des Diskurses sind, entsteht ein zusätzlicher Nutzen durch die Kombination der zwei Dimensionen. Das in (1) gezeigte Beispiel erhält durch die Inferenzregel in (4) die in Abbildung 1 gezeigte Annotation.

- (4) IF Kausalkonnektor gefunden AND

Kausalkonnektor von einer Partikel, der unabänderlichen Randbedingung markiert, gefolgt wird THEN
annotiere den start tag der Diskuseinheit mit
`<DiscRel="justification" CI="immutable_constraint">`

```
<discourse_unit id="17" DiscRel="justification" CI="immutable_constraint">
  <lexeme id="1" connector="causal">weil</lexeme>
  <lexeme id="2" particle="resignation_acceptance">halt</lexeme>
  <lexeme id="3">in</lexeme>
  <lexeme id="4">dem</lexeme>
  <lexeme id="5">Bereich</lexeme>
  <lexeme id="6">auch</lexeme>
  <lexeme id="7">die</lexeme>
  <lexeme id="8">meisten</lexeme>
  <lexeme id="9">Autos</lexeme>
  <lexeme id="10">unterwegs</lexeme>
  <lexeme id="11">sind</lexeme>
</discourse_unit>
```

Abbildung 1: Annotation Beispiel (1)

Im Gegensatz dazu annotiert die Annotationsregel in (5) die Kombination des Kausalkonnektors *da* mit der Modalpartikel *ja*; daraus folgt die Annotation des Beispiels aus (2) in Abbildung 2.

- (5) IF *da* ein Kausalkonnektor AND

da gefolgt von einer Partikel, der Gemeinsamkeit markiert THEN
annotiere den start tag der Diskuseinheit hit
`<discrel="justification" CI="common_ground">.`

```
<discourse_unit id="2" DiscRel="justification" CI="common_ground">
  <lexeme id="1" connector="causal">da</lexeme>
  <lexeme id="2">Sie</lexeme>
  <lexeme id="3" particle="common_ground">ja</lexeme>
  <lexeme id="4">gesagt</lexeme>
  <lexeme id="5">haben</lexeme>
</discourse_unit>
```

Abbildung 2: Annotation Beispiel (2)

Trotz der relativ geringen Größe des Korpus ist es schwer, übergreifende Muster in der Argumentativität über den Diskursverlauf hinaus auf einen Blick zu sehen, ohne den Blick auf einzelne Annotationen zu verlieren. Um diesem Hindernis zu begegnen, werden die Annotationen in einem Visualisierungssystem dargestellt, das die Muster zugänglicher macht. Dieses wird im Folgenden kurz erläutert.

4 Visualisierung von Argumentativität

Die Visualisierung linguistischer Muster hat in vielen Fällen gezeigt, dass diese Art der Informationsverarbeitung große Vorteile in der Erkennung von Mustern darstellt: bei theoretisch motivierten Phänomenen wie phonologischen Mustern (Mayer and Rohrdantz 2013) und Bedeutungswandel (Rohrdantz et al. 2011, 2012), bis hin zu Fragestellungen des maschinellen Lernens im Bereich des automatischen Clusterings (Lamprecht et al. 2013). Das Ziel der Visualisierung im vorliegenden Fall ist zum einen die Darstellung der Annotation der oben genannten Inferenzregeln; zum anderen kann die Verteilung der Annotation über den Diskursverlauf hinweg Aufschlüsse über den Grad der Deliberation liefern.

Abbildung 3 zeigt die Visualisierung eines Teils der S21-Mediationssitzung vom 4. November 2010; hierbei füllt jeder Satz eine eigene Zeile und jeder Sprechbeitrag wird von einem grauen Rechteck umrandet. Die gelben Balken markieren diejenigen Diskuseinheiten, die Kausalkonnektoren und damit Begründungen beinhalten.

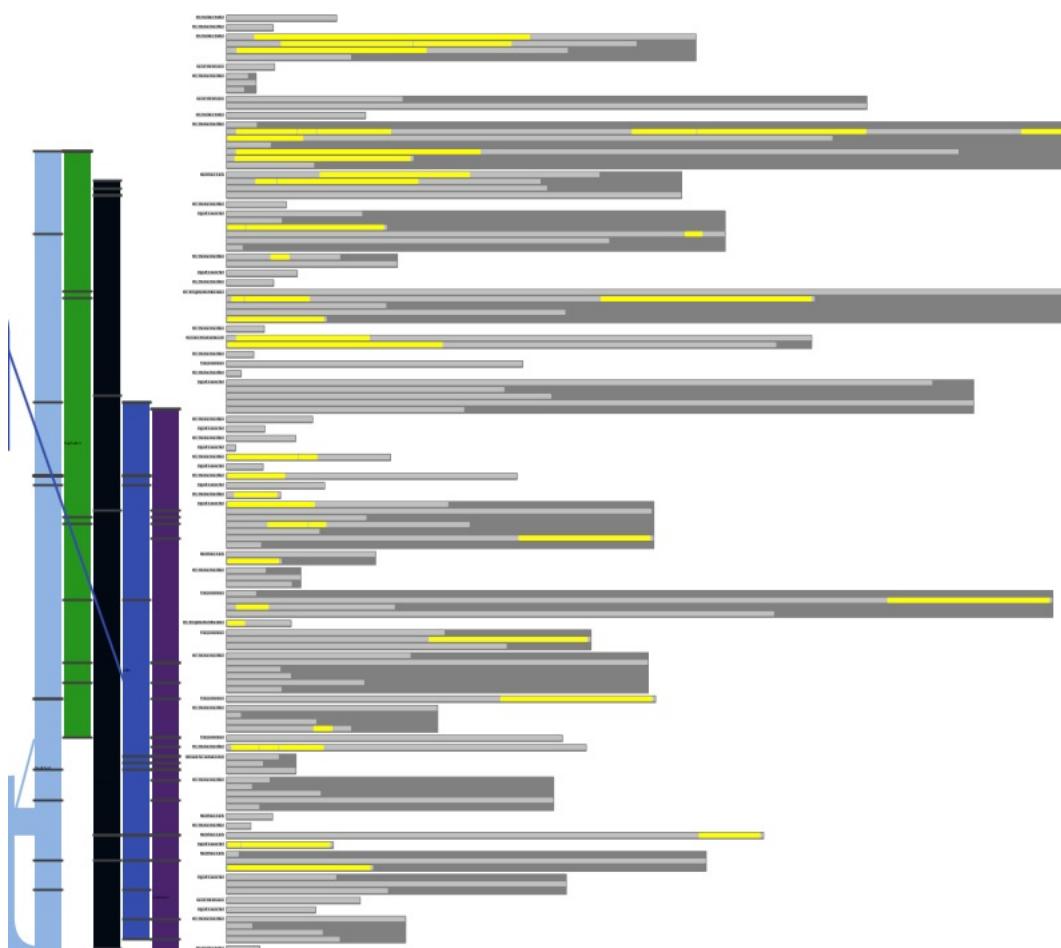


Abbildung 3: Visualisierung kausal begründeter Aussagen in der S21-Sitzung am 4. November 2010

Das entwickelte Werkzeug ist dahingehend interaktiv, als dass der Benutzer hinein- und herauszoomen und relevante Diskuseinheiten im Detail untersuchen kann, ohne die allgemeine Verteilung aus den Augen zu verlieren. Eine detaillierte Ansicht des Diskurses wird in Abbildung 4 gezeigt.

5 Zusammenfassung und weitere Fragestellungen

Dieses Papier präsentiert eine Methode der Operationalisierung des Begriffs der Deliberation durch Diskuskonnectoren und Modalpartikeln mit dem Ziel, die Mittel und Wege zu beleuchten, mit denen Argumente ausgetauscht und unter Sprechern und Zuhörern auf sie Bezug genommen wird. Durch die Verwendung eines Visualisierungsansatzes können die annotierten Datensätze über den gesamten Diskurs

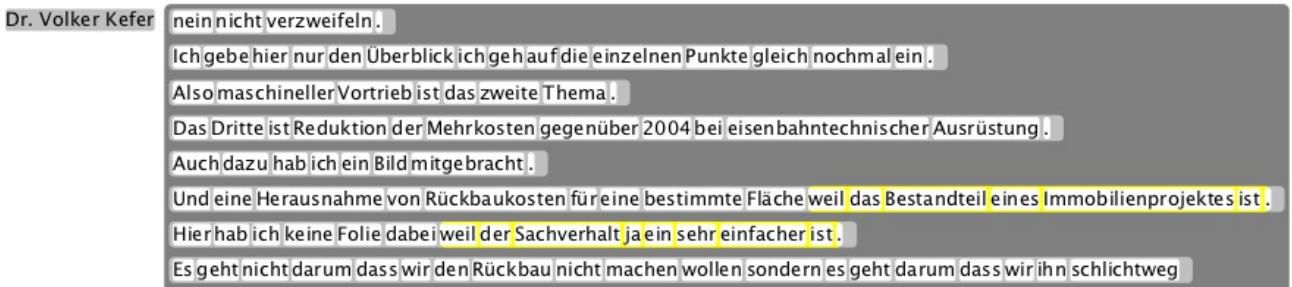


Abbildung 4: Detaillierte Visualisierung von begründenden Diskuseinheiten

hinweg inspiziert werden, was eine Interpretation der Rolle von Argumentativität im Meinungsbildungsprozess ermöglicht.

Künftige Ziele umfassen die Miteinbeziehung von zusätzlichen diskursrelevanten linguistischen Mustern sowie Multiword-Elementen. Bei steigender Anzahl der Annotationsebenen soll die Visualisierung entsprechend erweitert werden, sodass Interaktionen zwischen verschiedenen Ebenen aufgezeigt und tiefere Einsichten in Diskursstruktur und letztendlich Deliberation gewonnen werden können.

Literatur

- Bohman, James. 1996. *Public Deliberation: Pluralism, Complexity and Democracy*. Cambridge, MA: The MIT Press.
- Dryzek, John S. 1990. *Discursive Democracy: Politics, Policy, and Political Science*. Cambridge, MA: Cambridge University Press.
- Dryzek, John S. 2000. *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*. Oxford: Oxford University Press.
- Gutmann, Amy and Thompson, Dennis Frank. 1996. *Democracy and Disagreement. Why moral conflict cannot be avoided in politics, and what should be done about it*. Cambridge, MA: Harvard University Press.
- Habermas, Jürgen. 1981. *Theorie des kommunikativen Handelns*. Frankfurt am Main: Suhrkamp.
- Karagjosova, Elena. 2004. *The Meaning and Function of German Modal Particles*. Saarbrücken Dissertations in Computational Linguistics and Language Technology.
- Kwon, Min-Jae. 2005. *Modalpartikeln und Satzmodus: Untersuchungen zur Syntax, Semantik und Pragmatik deutscher Modalpartikeln*. Ph.D.thesis, LMU München.
- Lamprecht, Andreas, Hautli, Annette, Rohrdantz, Christian and Bögel, Tina. 2013. A Visual Analytics System for Cluster Exploration. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 109–114, Sofia, Bulgaria: Association for Computational Linguistics.
- Marcu, Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, Mass.
- Mayer, Thomas and Rohrdantz, Christian. 2013. PhonMatrix: Visualizing co-occurrence constraints in sounds. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Polanyi, Livia, Culy, Chris, van den Berg, Martin, Thione, Gian Lorenzo and Ahn, David. 2004. Sentential structure and discourse parsing. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 80–87.
- Prasad, Rashmi, Dinesh, Nikhil, Lee, Alan, Miltsakaki, Eleni, Robaldo, Livio, Joshi, Aravind and Webber, Bonnie. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961–2968.
- Rohrdantz, Christian, Hautli, Annette, Mayer, Thomas, Butt, Miriam, Keim, Daniel A. and Plank, Frans. 2011. Towards Tracking Semantic Change by Visual Analytics. In *Proceedings of the 49th Annual*

- Meeting of the Association for Computational Linguistics: Human Langauge Technologies (ACL-HLT '11): shortpapers*, pages 305–310, Portland, Oregon: Association for Computational Linguistics.
- Rohrdantz, Christian, Niekler, Andreas, Hautli, Annette, Butt, Miriam and Keim, Daniel. 2012. Lexical Semantics and Distribution of Suffixes — A Visual Analysis. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS and UNCLH*, pages 7–15.
- Zimmermann, Malte. 2011. Discourse Particles. In Paul Portner, Claudia Maienborn and Klaus von Heusinger (eds.), *Semantics (Handbücher zur Sprach- und Kommunikationswissenschaft)*, pages 2011–2038, Mouton de Gruyter.

Posterpräsentation: Digital Humanities Passau, 25.-28.03.2014

Slot: Beispiele für disziplinspezifische Anwendungen in der ganzen Breite der Geisteswissenschaften, sowohl in ihren objektbezogenen (Archäologie, Ur- und Frühgeschichte, Kunstgeschichte etc.) als auch in ihren textbezogenen Ausprägungen.

Die Schule von Salamanca. Eine digitale Quellensammlung und ein Wörterbuch ihrer juristisch-politischen Sprache

Ingo Caesar, Andreas Wagner

Abstract:

Das Projekt befasst sich mit dem einflussreichen juristisch-philosophisch-theologischen Diskurs der sog. Spanischen Spätscholastik, der im 16. und 17. Jahrhundert sein Zentrum auf der iberischen Halbinsel hatte. Die große Bedeutung seiner Autoren für das moderne Denken von Moral, Recht und Politik, ihre Wirkung auf verschiedenen Kontinenten und in unterschiedlichen akademischen Disziplinen ist in der modernen Forschung allgemein anerkannt, kann sich aber nur auf ein unzureichendes Bild des gesamten Diskurses stützen. Sowohl die spärliche Verfügbarkeit wichtiger Texte der damaligen Diskussionen als auch die disziplinäre Spezialisierung der heutigen Diskurse standen bislang weiter reichenden Einblicken entgegen. In diesen beiden Feldern soll das interdisziplinär aufgestellte Projekt einen substanzuellen Beitrag leisten.

Durch die Zusammenführung wichtiger Texte der Schule von Salamanca und ihre fachwissenschaftliche Erschließung eröffnet das vorzustellende Projekt einen für das moderne Denken von Ethik, Recht und Politik interdisziplinär relevanten Diskussionszusammenhang. Dabei werden eine aus 120 Werken bestehende Quellensammlung und ein zu erstellendes fachenzyklopädisches Wörterbuch unter Zuhilfenahme einer XML-Datenbank miteinander verknüpft und über eine Website verfügbar gemacht.

Das Poster soll einen Überblick über drei Themen geben:

(1) Workflow vom Bestandsnachweis der Quellen über die Sichtung, die Beschaffung, das Digitalisieren, das Double Keying, die TEI-Auszeichnung bis hin zur Präsentation in der Webanwendung.

Die Quellen (<http://salamanca.adwmainz.de/index.php?id=2134>) sind teils weltweit zerstreut. Ein kleiner aber wichtiger Teil konnte bisher nur in den USA oder in der chilenischen Nationalbibliothek nachgewiesen werden. Viele Exemplare lassen sich in Deutschland, vor allem in den Staatsbibliotheken München und Berlin, nachweisen. Ein großer Teil befindet sich in der Universitätsbibliothek von Salamanca und der spanischen Nationalbibliothek. Für die Digitalisierung gibt es mit mehreren dieser Einrichtungen bereits Kooperationsabkommen bzw. -gespräche. Erste Titel wurden bereits digitalisiert und gemeinsam mit dem Trierer Kompetenzzentrum und dessen Partner, dem chinesischen Erfassungsbüro TQY DoubleKey in Nanjing, sowie der digitalen Akademie in Mainz in TEI-XML-Volltexte überführt. Ein erster Prototyp der Datenbank und der Webanwendung steht vor der Fertigstellung.

(2) über die typographische und sprachliche Beschaffenheit des Quellenmaterials

Die meisten Texte aus dem 16. und 17. Jahrhundert sind in Latein verfasst, zu einem kleineren Teil in Spanisch. Neben einem oft mehrspaltigen Druckbild mit Summarien und Marginalglossen kommen häufig Abkürzungen, Brevigraphen und Allegationen vor. Ein großer Teil der Vorarbeit liegt in der Beschaffung gut lesbarer Vorlagen sowie der Erstellung einer Erfassungsanweisung für das Double-Keying, z.B. für die Darstellung von Sonderzeichen respektive das Einfügen von Platzhaltern. Auf deren Grundlage müssen solche Phänomene im Anschluss an die Retrodigitalisierung dann intellektuell aufgelöst werden. Neben diesen Auflösungen muss das sprachliche Material schließlich auch in weiteren Hinsichten mit diversen Suchfunktionen erschlossen werden. Dies stellt Anforderungen sowohl an die personellen Ressourcen und Kompetenzen des Projekts, muss aber auch in einer entsprechenden Infrastruktur durchgeführt werden und Aufnahme finden können.

(3) Die Webanwendung: Erschließung und Nutzen

Neben einem Einstieg über die Verzeichnisse von Werken, Autoren und Wörterbucheinträgen gibt es eine Volltextsuchfunktion, die es ermöglicht, in Überschriften, Summarien, oder nach normierten Orts- und Personennamen zu recherchieren. Darüber hinaus wird eine filternde Suche nach und innerhalb von Wörterbucheinträgen angeboten. Auch für die Suche innerhalb eines Werks wird eine Suche zur Verfügung gestellt.

Inhaltsverzeichnis, andere Indices im Text sowie eine Paginatorfunktion ermöglichen komfortables Navigieren im Werk. Ein Viewer gestattet die Anzeige der Images direkt neben dem Text. Aus dem Text heraus kann man zu den in den Belegstellen genannten Wörterbucheinträgen, sowie auf Übersichtsartikel zu den Autoren springen. Zitierfähige Hyperlinks werden für Abschnitte, Paragraphen, Seiten sowie Summarien angeboten. Ein weiteres Menü generiert die Anzeige im Text belegter normierter Personen und Orte und Belegstellen für die Wörterbucheinträge sowie der zitierten Literatur. Über einen Highlighting-Mechanismus lassen sich diese Belegstellen im Text hervorheben.

Damit bietet die Webanwendung auch ein Arbeitsinstrument für die Erstellung der Wörterbuchartikel. Wissenschaftlich relevante Stellen können recherchiert, zitiert und im Artikel ein direkter Verweis auf die Originalpassage hergestellt werden. Korpus und Wörterbuch werden direkt miteinander verlinkt: Belegstellen verweisen auf Wörterbuchartikel, Wörterbuchartikel auf relevante Stellen im Korpus.

Die Erstellung der Wörterbucheinträge und deren Verlinkung mit den Quellen sowie eine Suche, welche einen Wortformenabgleich berücksichtigt, sind einige der weiteren anstehenden Aufgaben und Herausforderungen.

Gegebenenfalls kann neben der Poster- auch eine Präsentation der Anwendung erfolgen.

Das auf achtzehn Jahre angelegte Projekt „Die Schule von Salamanca“ der Akademie der Wissenschaften und der Literatur, Mainz, startete im Februar 2013. Das Projektteam, das sich aus Mitarbeitenden des Instituts für Philosophie der Goethe-Universität Frankfurt und des Max-Planck-Instituts für europäische Rechtsgeschichte zusammensetzt, freut sich darauf, die Früchte des ersten Jahres zu präsentieren. Weitere Informationen unter <http://salamanca.adwmainz.de>

eCodicology

Mittelalterliche Handschriften als Gegenstand „archäologischer“ Forschung

Abstract für einen Vortrag auf der DHd-Tagung in Passau 2014.

Hannah Busch (Universität Trier)

Swati Chandra (Karlsruher Institut für Technologie)

Celia Krause (Technische Universität Darmstadt)

Oliver Schmid (Technische Universität Darmstadt)

Philipp Vanscheidt (Universität Trier / Technische Universität Darmstadt)

Der mittelalterliche Codex ist eine komplexe, in sich geschlossene Einheit, die aus unterschiedlichen Blickwinkeln heraus betrachtet werden kann: Text und Inhalt einer Handschrift werden von Sprach- und Literaturwissenschaftlern, Editionsphilologen, Historikern, Theologen und Religionswissenschaftlern untersucht. Mit der Schrift, ihrer Ausprägung und Entwicklung beschäftigen sich Paläographen. Die Buchmalerei wiederum ist Gegenstand der Kunstgeschichte. Für die Beschreibung der äußeren Merkmale des Codex hat sich eine eigene Wissenschaft herausgebildet, die in den 1940er Jahren erstmals mit dem Terminus *Kodikologie* belegt wurde und im romanischen Sprachraum (v. a. Frankreich und Italien) auch unter der Bezeichnung ‚Archäologie des Buches‘ bekannt ist. Kodikologische Forschung im engeren Sinne könnte man in ihrem Vorgehen teilweise mit Experimenteller Archäologie vergleichen: Der Codex wird wie ein archäologisches Artefakt in all seinen materiellen Bestandteilen studiert und beschrieben, um mehr über Arbeitstechniken und Abläufe in der mittelalterlichen Buchherstellung zu erfahren. Typische Untersuchungsgegenstände sind etwa Einband, Lagenstruktur, Linierungstechnik und Beschreibstoff.

QUANTITATIVE KODIKOLOGIE UND „MUSTERERKENNUNG“

Daneben ergeben sich kodikologische Fragestellungen, die weniger gut am Einzelobjekt überprüft werden können, etwa ob es bei der Herstellung bestimmte Normierungstendenzen und Entwicklungslinien gab. In diesem Fall untersucht man eine kritische Menge an Codices systematisch in Bezug auf einige charakteristische Parameter (z.B. Seitengröße, Anzahl der Textspalten, Zeilenzahl). Die festgehaltenen numerischen Werte lassen sich anschließend zueinander in Beziehung setzen und statistisch auswerten. Der unverwechselbare einmalige Charakter der Handschrift geht durch die quantitative Herangehensweise natürlich verloren, denn an die Stelle der Detailstudie am Einzelobjekt tritt die Untersuchung einer anonymen Masse ‚aus der Vogelperspektive‘. Allerdings lässt sich auf diese Weise mehr über den ‚gemeinen‘ Codex und über trendartige Veränderungen der Parameter im Laufe der Zeit herausfinden.

Dass sich der Computer für eine solche Aufgabe sinnvoll einsetzen lässt, kann das Projekt „eCodicology“ (<http://www.ecodicology.org>) demonstrieren. Seine Forschungsdaten bezieht das Projekt aus den rund 170.000 Codexseiten, die im Rahmen des Projekts „Virtuelles Skriptorium St. Matthias“ digitalisiert worden sind (<http://www.stmatthias.uni-trier.de>). Hierbei handelt es sich um ca. 450 Handschriften aus dem mittelalterlichen Bestand der Benediktinerabtei St. Matthias in Trier, die heute zum größten Teil in Trier selbst (Stadtbibliothek und Priesterseminar) aufbewahrt werden und in das 8. bis 16. Jahrhundert datieren. In „eCodicology“ sollen Layoutmerkmale auf den Digitalisaten automatisiert erkannt und extrahiert werden, die als konstituierend für die Gestaltung der Codexseiten gelten. Dazu zählen Seitenfläche, Schriftraum, Bildraum, freigelassener Raum, Textspalten, Zeilen und graphische Elemente. Geplant ist sowohl die Ermittlung von Ausdehnung und Anzahl der Elemente pro Seite als auch das Festhalten ihrer Position auf der Seite. Die neu gewonnenen Metadaten fließen in die Beschreibungen aus den früheren Handschriftenkatalogen ein. Für die einzelnen Merkmale sollen nachnutzbare Algorithmen entwickelt werden, durch deren Anwendung reproduzierbare Ergebnisse erzielt werden können. Mit Hilfe dieser Algorithmen können am Ende gezielte kodikologische Anfragen an den Bestand gerichtet werden. Eine statistische Auswertung soll schließlich Regelmäßigkeiten (Muster) bzw. Veränderungen innerhalb des Bestandes von St. Matthias aufzeigen; insbesondere sollen aussagekräftige Proportionen und Korrelationen zwischen einzelnen Konstanten gebildet werden. In diesem Zusammenhang ließe sich beispielsweise die Entwicklung der Verteilung von Bildraum und Textraum auf den Seiten verfolgen. Außerdem verspricht die Ermittlung von Layoutkonstellationen in Handschriften einen Erkenntnisgewinn in der Frage, ob es einen Zusammenhang zwischen dem Inhalt des Textes und der Anordnung der Seitenelemente gibt.

KODIKOLOGIE UND ARCHÄOLOGIE: DER KLEINSTE GEMEINSAME NENNER

Methoden der Mustererkennung spielen nicht nur in der Informationstechnik eine Rolle. Auch in vielen Geistes- und Kulturwissenschaften wird die Feststellung von Gemeinsamkeiten oder Gesetzmäßigkeiten bei Forschungsobjekten vorausgesetzt, seien es nun Texte, Bilder oder andere Untersuchungsgegenstände. Erst die Entdeckung und Klassifizierung von Mustern erlaubt es Geisteswissenschaftlern, Objekte zeitlich einzuordnen und Aussagen über kulturelle Bedeutungszusammenhänge zu treffen bzw. historische, politische, gesellschaftliche und religiöse Sinnzusammenhänge und Phänomene aufzuspüren. Über den rein praktischen Nutzen der automatischen Metadatenauszeichnung hinaus liefert das Projekt „eCodicology“ auch ein anschauliches Beispiel für eine gemeinsame Ausgangsbasis von kodikologischen und archäologischen Analysemethoden. Auch in der Klassischen Archäologie spielen Gestaltungsmuster eine Rolle: beispielsweise ermöglicht die Analyse von Haarlockenmotiven bei römischen Kaiserporträts ihre chronologische und typologische Einordnung. Wie im Projekt „eCodicology“ sind Maße und Zahlenverhältnisse auch in der Archäologie ein wichtiger Aspekt, z. B. für die Rekonstruktion eines griechischen Tempels. Im antiken Tempelbau orientierte man sich an grundlegenden Proportionen, vergleichbar etwa mit dem goldenen Schnitt bei der Aufteilung einer Buchseite. Aus diesen Überlegungen wird die Übertragbarkeit digitaler Methoden

ersichtlich, und zwar nicht nur vom Bereich der Informatik auf den Bereich der Geistes- und Kulturwissenschaften, sondern auch zwischen unterschiedlichen geisteswissenschaftlichen Disziplinen mit ähnlichen Arbeitsweisen.

ERSTE FALLSTUDIE (WORK IN PROGRESS)

Da die vorbereitenden technischen Schritte für die Prozessierung der Handschriftenscans wie Kalibrierung, Skalierung und Segmentierung noch nicht abgeschlossen sind, konnte eine automatisierte Merkmalsextraktion bisher nicht durchgeführt werden. Eine erste manuell durchgeführte Fallstudie mit einer repräsentativen Anzahl von Codices soll demonstrieren, wie die abschließende Analyse des Bestandes von St. Matthias aussehen könnte. Ziel wird es sein, die einzelnen Schritte einer statistischen Auswertung exemplarisch aufzuzeigen. In der Fallstudie soll ein Workflow erarbeitet werden, der auch bei der abschließenden Analyse automatisch erhobener Daten Anwendung finden kann. Die Untersuchung wird sich an wenigen grundlegenden Parametern orientieren, die unter anderem mit Hilfe der vorliegenden Metadaten aus dem Projekt „Virtuelles Skriptorium St. Matthias“ für eine größere Menge von Büchern relativ einfach zu erheben sind:

- *Seitengröße (Höhe x Breite)*: Höhen und Breiten der Seiten sollen stichprobenartig erhoben werden. Die manuellen Abmessungen werden am Bildschirm mit Hilfe der Software *Fiji* vorgenommen. Aus diesen Stichproben soll jeweils der Median pro Codex errechnet werden.
- *Datierung*
- *Gattung und/oder Inhalt des Textes*
- *Beschreibmaterial*
- *Anzahl der Textspalten auf den Seiten*

Diese Parameter können zueinander in Beziehung gesetzt werden und bilden die Basis für eine Clusteranalyse. Zum einen können Korrelationen zwischen zwei verschiedenen Parametern innerhalb einer Codexgruppe beobachtet werden; zum anderen können Vergleiche zwischen verschiedenen Codexgruppen mit unterschiedlichen Parametern angestellt werden. Größere Gruppen von Codices, die jeweils durch mindestens ein gemeinsames Merkmal definiert sind, können diachron analysiert werden. Im Idealfall lassen sich auf diese Weise ‚archäologische Schichten‘ im St. Mattheiser Handschriftenbestand erkennen oder mittels Regressionsanalyse entdeckte kausale Zusammenhänge nachweisen. Im Vortrag werden wir erste Ergebnisse der Analyse präsentieren und eruieren, inwieweit Methoden des Clustering und der Visualisierung gewinnbringend zur Anwendung kommen können.

New Technologies for Old Germanic. Resources and Research on Parallel Gospels in Older Continental Western Germanic

Christian Chiarcos
Gaye Detmolt
Maria Sukhareva

Jens Chobotský
Roland Mittmann

Goethe University Frankfurt am Main, Germany

December 16, 2013

We describe on-going efforts at the Goethe University Frankfurt on the study of older Continental Western Germanic languages, in particular, Old High German (OHG, antecessor of German) and Old Saxon (OS, antecessor of Low German and closely related to the antecessor of Dutch) and their relation to Old English (OE), Gothic, German and other Germanic languages as well as the relation of OHG and OS religious texts to their Latin sources. This line of research is conducted in the context of two larger efforts, the Old German Reference Corpus and the LOEWE cluster “Digital Humanities”, in collaboration with the Applied Computational Linguistics group at the Goethe-Universität Frankfurt.

The Old German Reference Corpus is a DFG-funded project that emerged from the Deutsch Diachron Digital (DDD) initiative, conducted in cooperation between HU Berlin, U Frankfurt and U Jena, and aims to provide a morphosyntactically annotated, exhaustive reference corpus of Old High German and Old Saxon. The LOEWE cluster “Digital Humanities”¹, funded through a programme of the State of Hessen, is a collaboration between U Frankfurt, TU Darmstadt and Freies Deutsches Hochstift Frankfurt aiming to develop methodologies and infrastructures to facilitate information-technological support of research in the humanities.

Here, we concentrate on biblical texts: These are available for a variety of modern and historical European languages and possess high-quality alignment (verses, segments), thus building up a valuable parallel resource for linguistic, philological and historical research questions, as well as for Natural Language Processing, whose methodologies for alignment and annotation projection can be used to support the analysis of these texts:

- The **Old German Reference Corpus** [4] provides a lexicon and an exhaustive corpus of older continental Western Germanic, i.e., Old Saxon (OS) and Old High German (OHG), comprising 650,000 tokens automatically enriched with morphological and morphosyntactic information drawn from existing glossaries which have been digitized by the project, complemented with manual annotations[3] and metadata and published via the ANNIS database [2].
- A **Historical Linguistic Database** was developed in LOEWE from a collection of etymological dictionaries for all Old Germanic languages (incl. OS, OHG, Old English, Gothic, Old Norse) as a relational data base providing user-friendly means of comparing etymologically related forms between historical dialects and their daughter languages, as well as a machine-readable view on these [5].
- Major texts in the corpus are the **gospel harmonies** associated with the names Heliand (OS), Tatian (OHG and Latin) and Otfrid (OHG). Although not direct translations of the

¹<http://www.digital-humanities-hessen.de>

Bible and hence not directly alignable with the gospel translations we have for Old English, Gothic, and later stages of English, German, Dutch and North Germanic, a section-level alignment has been manually extrapolated from the literature in a LOEWE project [5].

- This coarse-grained alignment is currently being refined to a **phrase-level alignment** using the linked lexical resources mentioned above as well as statistical models of systematic character correspondences like those applied by [1].
- On the basis of correspondences between historical and modern languages in parallel and quasi-parallel text, **statistical annotation projection** can be applied for the syntactic annotation of Older Germanic. So far, we conducted experiments on the joint projection of dependency syntax to Old English, Middle Icelandic and Early Modern High German corpora following the methodology of [6]. These indicate that projected annotations can serve as training data for mono- and cross-language parsing also for, e.g., OHG.
- These annotations can be applied, for example, to compare linguistic structures in OHG gospel harmonies and their Latin sources, thereby facilitating the research of a LOEWE project that currently uses statistical word alignment and existing *morphosyntactic* annotations only as the basis for a **qualitative, philological comparison** with the TreeAligner [7].

References

- [1] Marcel Bollmann. POS tagging for historical texts with sparse training data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW/ID-2013)*, pages 11–18, Sofia, Bulgaria, Aug 2013.
- [2] Christian Chiarcos, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitement Automatique des Langues (TAL)*, 49(2), 2008.
- [3] Sonja Linde and Roland Mittmann. Old German Reference Corpus. Digitizing the knowledge of the 19th century. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics / Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache*, Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus linguistics and Interdisciplinary perspectives on language (CLIP): 3, Tübingen, 2013. Narr.
- [4] Roland Mittmann. Digitalisierung historischer Glossare zur automatisierten Vorannotation von Textkorpora am Beispiel des Altdeutschen. *Journal for Language Technology and Computational Linguistics (JLCL)*, 27(2):39–52, 2013. Special issue *Altüberlieferte Sprachen als Gegenstand der Texttechnologie / Text Technological Mining of Ancient Languages*.
- [5] Timothy Blaine Price. Multi-faceted Alignment: Toward Automatic Detection of Textual Similarity in Gospel-derived Texts. In *Proceedings of Historical Corpora 2012*, Frankfurt, Dec 2012.
- [6] Kathrin Spreyer and Jonas Kuhn. Data-Driven Dependency Parsing of New Languages Using Incomplete and Noisy Training Data. In *Proceedings of CoNLL*, pages 12–20, Boulder, CO, Jun 2009.
- [7] Martin Volk, Joakim Lundborg, and Maël Mettler. A search tool for parallel treebanks. In *Proceedings of the 1st Linguistic Annotation Workshop (LAW-2007)*, pages 85–92, Prague, Czech Republic, Jun 2007.



**Digital Humanities – methodischer Brückenschlag oder „feindliche Übernahme“?
Chancen und Risiken der Begegnung zwischen Geisteswissenschaften und Informatik**

**Jahrestagung der Digital Humanities im deutschsprachigen Raum
Passau, 25.–28. März 2014**

Abstract (Poster)

**Zwischen Ton und Textgenese: Digital gestützte Verfahren zur kritischen Edition von
Operntexten in der *Online-Edition der Libretti zu Mozarts Opern***

Als Bestandteil einer multimedialen Gattung bezieht der Text einer Oper mehrere quellenkundliche, sprachwissenschaftliche und musikalische Ebenen mit ein, die in einem Druckmedium nur teilweise sinnvoll dargestellt werden können. Die *Online-Edition der Libretti zu Mozarts Opern*, die voraussichtlich im Januar 2014 im Rahmen der *Digitalen Mozart-Edition* erscheinen wird, bietet die Möglichkeit, all diese Dimensionen durch digital gestützte Verfahren nach wissenschaftlichen Kriterien synchron darzustellen.

Am Beispiel der Texte zu Wolfgang Amadé Mozarts *Le nozze di Figaro* wird gezeigt, wie Fassungen und Varianten parallel ediert werden, um einen direkten Vergleich der verschiedenen Quellenstränge zu ermöglichen. Durch die diplomatische Übertragung der Quellen mit Markierung aller Abweichungen zu den edierten Textfassungen lässt sich darüber hinaus jede einzelne editorische Entscheidung zum ersten Mal direkt zurückverfolgen. Als Verbindungsglied zur geplanten digitalen Edition des Notentextes wird schließlich mit einer metrischen Analyse des vertonten Textes ein Instrumentarium zur Erforschung der komplexen Beziehung zwischen dem metrischen Duktus der Textvorlage und dessen musikalischer Umsetzung zur Verfügung gestellt.

Ausstattung: Laptop mit Internet-Zugang und der Möglichkeit, Musik abzuspielen, Beamer.

CURRICULUM VITAE

Iacopo Cividini (Projektverantwortlicher), geboren 1975 in Bergamo (Italien), Studium der Musikwissenschaft und Geschichte an der Università degli Studi di Pavia (Italien), an der Johannes Gutenberg-Universität Mainz und an der University of Oregon (USA); Promotion in Musikwissenschaft an der Ludwig-Maximilians-Universität München im Jahre 2005 mit einer Arbeit über die Solokonzerte von Antonín Dvořák.

Wissenschaftlicher Mitarbeiter am DFG-Projekt *Bayerisches Musiker-Lexikon Online* an der LMU München von 2005 bis 2007 und an der *Digitalen Mozart-Edition* (DME) bei der Internationalen Stiftung Mozarteum Salzburg seit 2007.

Hauptforschungsgebiete: Instrumentalmusik des 19. Jahrhunderts, Bayerische Musikgeschichte, Libretti-Forschung, Philosophie der Aufklärung. Wichtigste Publikationen: *Die Solokonzerte von Antonín Dvořák. Eine Lösung der Konzertproblematik nach Beethoven*, 2007; Aufsätze zur Musik des 18. und 19. Jahrhunderts (u.a. Mozart, Brahms, Dvořák).

Digitale Projekte: *Bayerisches Musiker-Lexikon Online*, *Online-Katalog der Libretti zu Mozarts Opern*, *Online-Edition der Libretti zu Mozarts Opern* (geplante Veröffentlichung: Januar 2014).

Adriana De Feo, geboren 1980 in Salerno (Italien), studierte Kunst-, Musik- und Theaterwissenschaften an der Universität Bologna und schloss ihr Studium mit einer Arbeit über Mozart in Mailand ab. 2012 promovierte sie in Musikwissenschaft an der Universität Mozarteum Salzburg über das Thema *Mozarts Serenate im Spiegel der Gattungsentwicklung*.

Zu Ihren Forschungsschwerpunkten zählen die Huldigungsoper des 17. und 18. Jahrhunderts und die Libretti-Forschung. Aufsätze zur Musikdramaturgie des Barock und der Klassik.

Als Wissenschaftliche Mitarbeiterin der Stiftung Mozarteum Salzburg (seit 2009) erarbeitet sie im Rahmen der *Digitalen Mozart-Edition* (DME) die Online-Edition der Libretti und den Online-Katalog der Libretti zu Mozarts Opern.

Franz Kelnreiter

Studium der Musikwissenschaft und Romanistik (Französisch) an der Paris-Lodron-Universität sowie der Kirchenmusik an der Hochschule Mozarteum in Salzburg. Seit 1994 an der Internationalen Stiftung Mozarteum, zunächst als Leiter der Mozart Ton- und Filmsammlung, seit 2002 Mitarbeiter an der *Digitalen Mozart-Edition* und Leiter des IT-Bereichs. Technische Projektbetreuung.

Salzburg, 12.12.2013

Iacopo Cividini, Adriana De Feo, Franz, Kelnreiter

Albertina – Sammlungen online: eine digitale Ressource und ihre Nutzung

Abstract zu einem Vortrag

Jahrestagung der Digital Humanities im deutschsprachigen Raum, März 2014, Universität Passau
Regina Doppelbauer, Albertina Wien

Der Vortrag ist ein Praxisbericht aus dem Museum und geht von der subjektiven Wahrnehmung aus, dass zwischen den Institutionen, die Daten zur Verfügung stellen und den Nutzern aus den digital humanities kaum Dialog und Kommunikationswege bestehen. Zwischen der Praxis der Materialerschließung und den Anliegen der digital humanities scheint sich eine schwer überbrückbare Kluft aufzutun, für die verschiedene Interessen, unterschiedliche Sprachen, Terminologien sowie Diskursebenen verantwortlich sind.

Der Beitrag thematisiert, ob der Versuch einer Überbrückung für beide Seiten fruchtbar sein kann. Er stellt Fragen nach der prinzipiellen Wahrnehmung einer bewahrenden Institution durch die digital humanities, nach der Einbindung der bereitgestellten digitalen Ressourcen in deren Recherche- und Nutzungsstrategien und nach den Möglichkeiten des Feedbacks. Denn: Das durch die Museen liberalisierte Wissen stellt keine *l'art pour l'art* – Übung dar, sondern soll so gut wie möglich von der wissenschaftlichen community – und auch von anderen Öffentlichkeiten - aufgenommen und genutzt werden.

Um Verständnis für die Anforderungen, Arbeitsabläufe und kuratorischen Aspekte des Museums und die besondere Ausgangssituation der Albertina-Datenbanken herzustellen, werden das Haus und seine Digitalisierungs- und online-Strategie einleitend vorgestellt.

Die Albertina

Die Albertina - eine der größten und bedeutendsten grafischen Sammlungen weltweit - war in der Museumslandschaft bis zum Jahr 2000 als „Graphische Sammlung Albertina Wien“ verankert. Unter dem im Jahr 2000 neu angetretenen Direktor Klaus Albrecht Schröder erfolgten zwischen 2000 und 2003 tiefgreifende Umbau-, Renovierungs- und Erweiterungsarbeiten. Restaurierwerkstätten, Bibliothek, Fotoatelier, Studiensäle sowie ein automatisches Hochregallager für rund 1 Million Kunstwerke wurden im Erdkern der Basteianlagen, auf denen das historische Palais steht, untergebracht. Die Ausstellungsflächen wurden von 250 m² auf 3.800 m² vergrößert.

In der öffentlichen Wahrnehmung:

2003 wurde die „Albertina“ – neu positioniert als internationales Ausstellungshaus - wieder eröffnet. Die Albertina präsentiert seither neben einer permanenten Schausammlung von Gemälden und Skulpturen („Von Monet bis Picasso“, „Albertina contemporary“) parallel mehrere Ausstellungen (Alte Meister, Klassiker der Moderne, Fotografie, Personalen zeitgenössischer Künstler).

„Backstage“: Die Sammlungen der Albertina

Gesamt:	1,120.000 Objekte
Gemälde und Skulpturen:	600 Objekte
Grafische Sammlung:	50.000 Zeichnungen 900.000 Druckgraphiken
Architektursammlung:	50.000 Pläne und Skizzen
Fotosammlung:	100.000 Objekte
Plakatsammlung:	20.000 Objekte

Digitalisierung und online-Datenbank

Seit 1999 werden die Bestände in der TMS-Datenbank (The Museum Systems/Gallery Systems) digitalisiert: 250.000 Datensätze, davon 190.000 mit Image, sind gegenwärtig intern verfügbar. TMS wird für zahlreiche Arbeitsprozesse verwendet (Dokumentation der Sammlung, Leihgabenmanagement, Restaurierungsabteilung).

Online-Datenbank:

2007 erste online-Präsentation von rund 25.000 Objekten mit flachen Daten;
2012 offensive Erweiterung der Anzahl online zugänglicher Objekte mit tiefer Datenerreichbarkeit:
Datenkontrolle, Anreicherung mit Katalogtexten;
Relaunch mit neuer Technologie (CIT/Den Haag): unter einem Portal sind drei Datenbanken der Albertina abrufbar (Bilddatenbank, Biobibliographie zur Fotografie in Österreich, Bibliothek der Albertina).

2014: 50.000 Objekte mit Images online recherchierbar, 9.000 davon mit vertiefenden Texten (<http://sammlungenonline.albertina.at/1.8/Default.aspx> - Achtung, gegenwärtig noch Beta-Version, live ab Jänner 2014).

Nutzung

Die Sammlungsobjekte der Albertina (zum allergrößten Teil Arbeiten auf Papier!) können aus konservatorischen Gründen nicht permanent ausgestellt werden. Meist werden nur kleine Ausschnitte aus den prominentesten Beständen (Albrecht Dürer, P.P. Rubens, Rembrandt, Klimt, Schiele) in Ausstellungen präsentiert und damit durch print-Kataloge erschlossen. Die online-Datenbank macht daher Verborgenes sichtbar und stellt mit ihren reichen Daten (Metadaten, vertiefende Texte, Images) der Forschung gutes Material zur Verfügung. Diese soll unsere Objekte leicht finden können – und wir wollen gefunden werden!

Wir kennen aus dem Feedback und aus der Besucherstatistik in groben Zügen die gegenwärtigen Nutzer der online-Datenbank: Es handelt sich um kunsthistorisch interessiertes Publikum und um Wissenschaftler. Die Strategien, weitere und breitere Nutzerschichten zu erschließen, umfassen mehrere Ebenen: Die gute Sichtbarkeit innerhalb der Museums-Website, ein an aktuelle Seherwartungen angepasstes Layout und freundliche Angebote für Netz-Flaneure adressieren diejenigen, die die Website direkt aufrufen. Einbindungen in übergeordnete Portale können zu einer diffusen Multiplikation (Europeana) oder zu einer gezielten Ansprache von Usern (Prometheus; projektiertes Graphikportal von Foto Marburg) genutzt werden.

Doch was muss hinsichtlich der Nutzung durch die digital humanities als Voraussetzung gewusst werden, um adäquat reagieren und funktionierende Zugänge legen zu können?

Datenverbesserung: Die online verfügbaren Daten der Albertina sind mittels deep links und permalinks für Google offen und auffindbar. Wie sollen die Daten darüber hinaus modelliert, annotiert, angereichert werden, um in big data/im semantic web an relevanter Stelle auffindbar zu sein?

In der Kunstgeschichte würde sich beispielsweise die Aufgabe stellen, die Ikonographie mittels Iconclass eindeutig festzulegen. Doch dies stellt einen Arbeitsaufwand dar, der bei vielen Tausend Objekten kaum zu leisten ist. Wäre er denn pro futuro überhaupt wichtig oder ist er bereits obsolet? Wer würde die genaue Iconclass-Nummerierung bzw. -Terminologie (z.B.

<http://www.iconclass.org/rkd/11F/>) zum Angelpunkt seiner Recherche machen? Gehen die Suchwege der Forscher nicht vielmehr – wie die der meisten User – über Google? Die Google-Algorithmen

verlangen für die Darstellung an relevanter Stelle wiederum ganz andere Qualitäts- bzw. „Erfolgs“kriterien!

An welchem Ort stellt sich die digitale Ressource „Bilddatenbank“ im Netzwerk der digital humanities dar? Welche Suchwege beschreiten bzw. forcieren die digital humanities, welche Plattformen kreieren sie neu? Die Nutzung von Google scheidet die Geister - wird ein europäisches Forschungsinfrastrukturprojekt wie Dariah eine relevante Alternative darstellen können?

Nutzerwünsche: Bezüglich der wissenschaftlichen Nutzer hat sich subjektiv ein Bild entwickelt, das diese – holzschnitthaft – nach zwei Gruppen unterscheidet: den „klassischen“ Kunsthistorikern, die oft noch in großer Nähe zum physischen Objekt, und den digitalen Geisteswissenschaftlern, die auf der Ebene übergeordneter Diskurse und Strukturen operieren. Die Etablierung einer Feedbackkultur würde es den Erstellern wissenschaftlicher Bilddatenbanken ermöglichen, ihre Daten noch gezielter auszustalten und den Nutzern entgegen zu kommen. Im Rahmen einer erst jüngst abgehaltenen Tagung in Wien

(http://www.mediathek.at/ueber_die_mediathek/aktuelles/details/article/authentisch-im-netz-einladung-zur-wissenschaftlichen-tagung/) wurde der Wunsch nach direktem Feedback auch von VertreterInnen einer großen Zeitungs- und einer audiovisuellen Datenbank geäußert.

Direktes Feedback würde z.B. bedeuten, die Qualität, mit der eine Quelle im Web erschlossen wird, hinsichtlich der technischen und/oder inhaltlichen Ebene zu kommentieren und Vorschläge zu einer Verbesserung anzuschließen. Ich denke dabei nicht an dezidiert kollaborative Modelle, die ein gemeinsames Forschen meinen, sondern an schlichte Kommentare zum Gebrauch. Der Button „Möchten Sie uns etwas mitteilen?“ ist schließlich bei den meisten Datenbanken vorhanden!

Diejenigen, die in Museen mit dem Material selbst arbeiten, sind oft – trotz fachwissenschaftlichen Studiums – von den laufenden wissenschaftlichen Diskursen weit entfernt (KuratorInnen sind davon auszunehmen). Das schlichte Interesse, entlang welcher Fragestellungen sich Geisteswissenschaften heute bewegen, ist zu diffus, um mit Antworten rechnen zu dürfen. Dennoch: Stimmt der Eindruck, dass gegenwärtiges (kunst/historisches) Forschen weniger vom Material, sondern von Hypothesen ausgeht, die sich ihr Material erst suchen?

Interessant wäre auch zu erfahren, welche Einstellungen und Erwartungen sich hinsichtlich des Downloads und der Weiterverwendung von Bildmaterial (Urheberrecht und creative commons) herausbilden.

Die Entwicklung eines Kommunikationsnetzes nach vielen Seiten hin und das Pflegen einer entsprechenden Kultur könnten helfen, dem weiteren Auseinanderdriften von musealer Grundlagenbereitstellung und verarbeitender Wissenschaft entgegenzuwirken. Die Interessenslagen und Terminologien sind denkbar verschieden und das Bemühen um wechselseitiges Verstehen wird wohl Übersetzungsleistung und Vermittlungsebenen benötigen. Wird deren Notwendigkeit so auch von der Wissenschaft/den digital humanities gesehen und kann ein Wille in diese Richtung gebildet werden?

Nicht zuletzt geht es auch darum, die unterschiedlichen Wissensstände und Voraussetzungen mehrerer Generationen – digital natives und digital immigrants – anzuerkennen und, wenn möglich, zu harmonisieren.

Digital Humanities – methodischer Brückenschlag oder „feindliche Übernahme“?

Chancen und Risiken der Begegnung zwischen Geisteswissenschaften und Informatik

Jahrestagung der Digital Humanities im deutschsprachigen Raum

25.-28. März 2014, Universität Passau

PECHA KUCHA VIRTUELLE REKONSTRUKTION

ALLGEMEINE STANDARDS, METHODIK UND DOKUMENTATION

Panel-Organisatoren: Piotr Kuroczyński (Herder-Institut Marburg) und Mieke Pfarr-Harfst (TU Darmstadt)

Das Panel beschäftigt sich mit dem Thema der *virtuellen Rekonstruktion* von gebautem kulturellen Erbe – einem interdisziplinären Arbeitsgebiet an der Schnittstelle zwischen Archäologie, Architektur, Bau- und Kunstgeschichte, Geschichte, Soziologie und Informatik.

Die *virtuelle Rekonstruktion* gewinnt, bedingt durch den *spatial- und iconic-turn*, immer mehr an Bedeutung für eine räumlich-visuelle Wissensordnung in Zeiten der Allgegenwärtigkeit der digitalen Information und ihrer Georeferenzierung.

Die medientechnologische Entwicklung der letzten 25 Jahre zeugt von einem Siegeszug der *virtuellen Rekonstruktion*. Die Verbreitung und Popularität der Technologie u. a. im Kontext von Lehre und Forschung sowie der Vermittlung von Wissenständen ist ein bekannter Ausdruck davon. Das Arbeitsfeld wird von Architekten, Bauhistorikern, Archäologen und Informatikern geprägt, die fachbedingt als Pioniere auf dem Gebiet der geschichtswissenschaftlichen Rekonstruktion in Forschung und Vermittlung anzusehen sind. Die Kunstgeschichte als klassisches Teilgebiet der Geisteswissenschaften, die den neuen Medien bisher eher kritisch gegenüberstand, öffnet sich in der letzten Dekade den Informations- und Kommunikationstechnologien immer mehr. So ist mit den *Digital Humanities* ein eigenes Forschungsfeld entstanden, in dem die *virtuelle Rekonstruktion* anzusiedeln ist.

Die bisher meist nur anwendungsbezogenen *virtuellen Rekonstruktionen* werden so selbst zum Forschungsgegenstand, um ihre Potentiale als Forschungswerkzeug in vollem Umfang ausnutzen und eine fundierte einheitliche Basis schaffen zu können. Einige ausgewählte Fragestellungen sind: Wie wissenschaftlich sind *virtuelle Rekonstruktionen* als Forschungs- und Vermittlungsmethode? Wie ist es möglich, das in den digitalen Modellen vorhandene Wissen im Sinne einer wissenschaftlichen Dokumentation nachhaltig zu sichern und nachzuweisen? Welche Arbeitsmethodik, Datenformate, Datenbanken sind hierfür nötig auch in Hinblick auf eine Langzeitarchivierung und Verfügbarkeit? Sind webbasierte *virtuelle Forschungsumgebungen* die Arbeitsplattform der Zukunft? Wie müssen sie aufgebaut werden, um eine hohe Interoperabilität der Datensätze, einen intuitiven Zugang und leichte Editierbarkeit zu gewährleisten? Wie sind *virtuelle Rekonstruktionen* in der Vermittlung für eine breitere Nutzergruppe zu gestalten?

Die *virtuellen Rekonstruktionen* werden bis heute von den Geisteswissenschaften kritisch bewertet, da man keine gültigen Antworten für eine „kritische Computer-Visualisierung“ (H. Günther, 2001) geliefert hat. Gleichzeitig steigt die Zahl an Rekonstruktionsprojekten stetig an, institutionsübergreifende Maßstäbe oder

verbindliche Standards hinsichtlich der Wissenschaftlichkeit fehlen. Für deren Etablierung mangelt es vor allem an interdisziplinären Langzeitkooperationen und einer nachhaltigen Vernetzung der Community. Ein immenser und nicht länger vertretbarer Wissens- und Ressourcenverlust innerhalb der europäischen Lehr- und Forschungslandschaft ist die Folge. Die Zahl der Projekte, die in ein paar Jahren noch lesbar und nachvollziehbar, d. h. im wissenschaftlichen Sinne verwendbar sind, ist schwindend gering.

Zwar sind Ideen und Forderungen wie die *Architectura Virtualis* (M. Koob, 1995) oder die *Londoner Charta* (R. Beacham, H. Denard, F. Niccolucci, 2006) aufgestellt und anerkannt, aber ein konkret umgesetzter Ansatz ist bis dato nicht vorhanden. In den letzten fünf Jahren sind vermehrt Verbundprojekte entstanden, welche die *virtuelle Rekonstruktion* als Werkzeug der Wissenschaft einsetzen und erforschen. Doch ist auf diesem Forschungsgebiet ein über die Verbundprojekte hinausgehender Konsens, der eine breitere Gültigkeit erlangt, nicht erkennbar.

In Anbetracht der Notwendigkeit einer breiteren Vernetzung und langfristiger Kooperation ist es unabdingbar, ein interdisziplinäres Arbeitsgebiet der *virtuellen Rekonstruktion* innerhalb der *Digital Humanities*, zu formieren und nachhaltig zu etablieren. Ziel ist, eine Austauschplattform zu errichten, um gemeinsam den Weg hin zu allgemeinen Standards bei der *virtuellen Rekonstruktion* vorzubereiten. Der akute Handlungsbedarf eröffnet zugleich die Möglichkeit, sich eine Vorreiterrolle auf dem Gebiet der Dokumentation und Langzeitverfügbarkeit von digitalen Wissensräumen im Sinne der *Charta zur Bewahrung des digitalen Kulturerbes* (UNESCO, 2003) zu sichern.

Die gewählte Vortragsform, das Pecha-Kucha-Format, soll der Vielschichtigkeit des Forschungsfeldes gerecht werden und in Form kurzer prägnanter Statements die aktuelle Forschungslandschaft im Bereich der *virtuellen Rekonstruktion* beleuchten. Pecha-Kucha ist eine bewährte Präsentationstechnik, die einen Ideenaustausch in der anschließenden Diskussion fördert. Der Vortragende muss durch eine limitierte Anzahl von 20 Bildfolien und eine vorgegebene Zeit von 20 Sekunden je Folie seinen Vortrag und die damit verbundene Botschaft auf den Punkt bringen. Die Statements jedes einzelnen werden parallel während des Vortrags auf einem zweiten Projektor fixiert, die anschließend als Reminder stehen bleiben und die Grundlage für die Diskussion bilden. Der Fokus dieses Panels liegt auf der wissenschaftlichen Diskussion und dem Austausch der Partizipierenden.

Sander Münster (TU Dresden) beleuchtet im Rahmen seines Dissertationsvorhabens Kooperationsprozesse, -phänomene und -strategien bei der interdisziplinären Erstellung von virtuellen 3D-Rekonstruktionen aus sozialwissenschaftlicher Perspektive. Anhand einer bibliometrischen Analyse von Publikationen zu geschichtswissenschaftlicher 3D-Modellierung soll im Kurzvortrag untersucht werden, wie verbreitet einschlägige methodische Standards wie beispielsweise die bereits benannte *Londoner Charta* oder Metadatenstandards wie CIDOC-CRM in der Projektpraxis sind.

Darüber hinaus soll ein besonderes Augenmerk darauf geworfen werden, ob und wie sich derartige Standards in der Projektpraxis manifestieren. Dabei soll der Vortrag anhand von Fallbeispielen aufzeigen, welche Alltagsstrategien sich bezüglich Zusammenarbeit, Sicherstellung von Wissenschaftlichkeit und Arbeitsvorgehen im Verlauf von 3D-Rekonstruktionsprojekten herausbilden und welche Anknüpfungspunkte sich daraus für eine zielgerichtete Entwicklung von Methoden und Werkzeugen ergeben.

Mieke Pfarr-Harfst (TU Darmstadt) hat über das Thema *Dokumentationssystem für Digitale Rekonstruktionen* promoviert, um das in den dreidimensionalen Modellen vorhanden Wissen nachhaltig zu sichern.

Ausgehend von den Potentialen *Digitaler Rekonstruktionen* als Forschungsmethode in einem interdisziplinären Rekonstruktionsprozess, der zur Wissensfusionierung und -generierung führt, wird das

Dokumentationssystem im Rahmen des Panels erläutert und zur Diskussion gestellt. Hierbei werden die vier aufeinander aufbauenden Ebenen des Gesamtsystems kurz dargestellt.

Ebene 1 – Projekthintergrund: Informationen zu Entstehungsjahr, Projektbeteiligten, Stand der Forschung

Ebene 2 – Projektkontext: Der kulturelle, historische und bauhistorische Kontext des Projektes

Ebene 3 – Systematik: Festlegen der individuellen Regelwerke wie Nomenklatur, Klassifizierungen

Ebene 4 – Nachweisebene: Direkte Zuordnung von Rekonstruktionsobjekt (Gebäude, Struktur) zu den Quellen und dem Prozess mittels textbasierten Baubeschreibungen, Quellen- und Methodenkatalogen

Frank Henze und Henning Burwitz (BTU Cottbus) resümieren langjährige Erfahrungen mit virtuellen Rekonstruktionen aus verschiedenen bauhistorischen und archäologischen Projekten. Aus den jeweiligen projekt- und fachspezifischen Anforderungen an die Modellierung, Analyse und Präsentation der Rekonstruktionen werden grundlegende Anforderungen an den Arbeitsablauf und die virtuelle Arbeitsumgebung abgeleitet, vor allem aber Potentiale und Risiken der technischen Möglichkeiten kritisch diskutiert.

Darauf aufbauend wird ein Einblick in die aktuelle Entwicklung des webbasierten Informations- und Dokumentationssystems *OpenInfRA* gegeben, in das umfangreiche Funktionen zur Verarbeitung, Analyse und Präsentation von Geometriedaten integriert werden.

Der Vortrag erläutert im speziellen die Konzeption und Implementierung eines 3D WebGIS. Neben der Visualisierung im Browser spielen dabei die Datenhaltung in Geo-Datenbanken sowie die Verwendung standardisierter Dienste und Formate zum Austausch und zur Abfrage von 3D-Geometrien eine wesentliche Rolle.

Es wird gezeigt, dass die Möglichkeiten einer umfassenden Attributierung von 3D-Objekten sowie raumzeitliche Abfragen in *virtuellen Rekonstruktionen* das Potential haben, die wissenschaftliche Arbeit methodisch zu erweitern.

Piotr Kuroczyński (Herder-Institut Marburg) stellt am Beispiel des aktuellen Verbundprojektes *Virtuelle Rekonstruktionen in transnationalen Forschungsumgebungen – Das Portal: Schlösser und Parkanlagen im ehemaligen Ostpreußen* die inhaltlich bedingten Anforderungen sowie die methodischen und technischen Lösungsansätze vor.

Der Kurzvortrag thematisiert die Rückverfolgung der einzelnen Arbeitsschritte und Entscheidungsprozesse sowie der ihnen zugrunde liegenden Quellen. Darüber hinaus wird die Software-Architektur einer virtuellen Forschungsumgebung angerissen und ein Einblick in den Front- und Backend des Web-Portals gegeben. Der Beitrag geht auf die Integration von semantischen 3D-Objekten im WWW, das Regelwerk CIDOC CRM/LIDO, die Datenformate XML/RDF und OBJ/COLLADA/JSON, sowie die WebGL-Technologie ein und möchte eine Debatte über praktikable Methoden und Techniken anregen.

Markus Wacker (HTW Dresden) stellt anhand der Projekte zum *Dresdner Zwinger* und *Muristan* in Jerusalem die computergestützten Möglichkeiten vor, mit denen vielschichtige, multimediale, objektbezogene Quellen für die Recherche und Forschung nutzbar gemacht werden.

Der Beitrag thematisiert die Suche nach geeigneten Metaphern für die Visualisierung und Dokumentation. Im Vordergrund stehen informationstechnische Lösungsansätze, adäquate Nutzerschnittstellen und praktikable Verknüpfungsmöglichkeiten der Informationsquellen, mit der dahinter stehenden Datenbankstruktur und Datenformaten, sowie die Informationsvisualisierung der Datenbank.

Pilotphase Handschriftendigitalisierung

Im Juni 2013 genehmigte die DFG eine zweijährige Pilotphase zur Digitalisierung mittelalterlicher Handschriften an den deutschen Handschriftenzentren.

Auf Basis der praktischen Erfahrungen aus sieben Digitalisierungsprojekten, die an fünf Bibliotheken durchgeführt werden, soll ein nationaler Masterplan erarbeitet werden, der bei positiver Begutachtung zur Grundlage einer zukünftigen DFG-Förderlinie wird. Die projektübergreifenden Organisations- und Koordinationsarbeiten sind an der Bayerischen Staatsbibliothek angesiedelt.

Neben Priorisierungsfragen steht die Entwicklung einer Infrastruktur im Mittelpunkt, die bestandshaltenden Institutionen in ganz Deutschland die Digitalisierung relevanter Bestände auf hohem, den aktuellen Standards entsprechendem Niveau ermöglicht. Zentraler Zugriffspunkt für Digitalisate wie für zugehörige Meta-, Erschließungs- und Strukturdaten ist das deutsche Handschriftenportal *Manuscripta Mediaevalia*, das in Zusammenarbeit mit dem Bildarchiv Foto Marburg fortwährend weiterentwickelt wird.

Mithilfe eines Posters möchten wir über den aktuellen Projektstand informieren und zur Diskussion einladen. Zu den folgenden Themenbereichen sollen Informationen graphisch aufbereitet werden:

Hintergründe und Ziele

Um im Bereich der Digitalisierung möglichst effektive Förderentscheidungen treffen zu können, hat die DFG in der Vergangenheit bereits mehrmals die Entwicklung eines übergreifenden Projektplans in Auftrag gegeben (sog. „Masterplan“ oder „Roadmap“). In einem derartigen Plan werden Verfahrensstandards festgelegt, Prioritäten gesetzt und Kosten kalkuliert; er wird in der Regel im Rahmen einer Pilotphase entwickelt, die mehrere koordinierte Einzelprojekte umfasst.

Das nun bewilligte Vorhaben für den Bereich der Handschriftendigitalisierung besteht aus sieben Einzelanträgen und einem flankierenden Rahmenantrag. Letzterer beinhaltet die Koordination der Einzelprojekte sowie die technische Weiterentwicklung des Portals *Manuscripta Mediaevalia*.

Projektpartner

- Staatsbibliothek zu Berlin
- Universitätsbibliothek Leipzig
- Bayerische Staatsbibliothek München
- Württembergische Landesbibliothek Stuttgart
- Herzog-August-Bibliothek Wolfenbüttel
- Bildarchiv Foto Marburg

Fallgruppen und Einzelprojekte

Die in den Einzelprojekten zu digitalisierenden Bestände unterscheiden sich bezüglich ihres Erschließungsgrads und der Schwierigkeit der Digitalisierung; sie wurden als repräsentativ für die Handschriftenüberlieferung in Deutschland ausgewählt. Es wurden fünf Fallgruppen definiert:

- *Fallgruppe 1:* Begleitende Digitalisierung bei laufenden Tiefenerschließungsprojekten
 - Projekt 1: Digitalisierung lateinischer Handschriften aus dem ehemaligen Benediktinerkloster St. Emmeram in Regensburg (München)
 - Projekt 2: Digitalisierungskomponente zum Projekt „Erschließung von Kleinsammlungen mittelalterlicher Handschriften in Sachsen und dem Leipziger Umland“ (Leipzig)
- *Fallgruppe 2:* Digitalisierung gut erschlossener Bestände
 - Projekt 3: Digitalisierung von Handschriften der Ratsbücherei Lüneburg (Wolfsbüttel)
 - Projekt 4: Digitalisierung mittelalterlicher deutscher Pergamenthandschriften aus dem Signaturenbereich Cgm 1–200 (München)
 - Projekt 5: Digitalisierung von Handschriften des Fonds Codices biblii in Folio (Stuttgart)
- *Fallgruppe 3:* Digitalisierung ungenügend erschlossener Bestände
 - Projekt 6: Digitalisierung von Handschriften der Signaturengruppe „Manuscripta germanica“ unter Nutzung aktualisierter historischer Kurzkataloge (Berlin)
 - Projekt 7: Bestandslistenerfassung und Digitalisierung von Handschriften ohne publizierten Nachweis sowie von zehn stark nachgefragten, aber nur mit deutlich erhöhtem Aufwand zu digitalisierenden Handschriften (Leipzig)
- *Fallgruppe 4:* Digitalisierung aufgrund aktueller Forschungsinteressen
- *Fallgruppe 5:* Digitalisierung mit deutlich erhöhtem Aufwand (Handschriften in jedem der Einzelprojekte anteilig enthalten)

Ein für die Fallgruppe 4 beantragtes Digitalisierungsprojekt wurde nicht bewilligt. Damit die Perspektive der Forschung dennoch Eingang in den Masterplan findet, ist die Auswertung anderer Forschungsprojekte mit Digitalisierungskomponente geplant.

Fallgruppe 5 enthält Stücke, deren Digitalisierung aufgrund ihrer Materialität oder ihres konservatorischen Zustands besonders schwierig ist. Mithilfe der so gewonnenen Daten wird sich der mit der Digitalisierung verbundene Aufwand präzise beziffern lassen.

Masterplan

Die Koordination der Einzelprojekte übernimmt die Bayerische Staatsbibliothek. Wesentlicher Bestandteil der Koordinationsarbeiten ist die Auswertung der eingesetzten Verfahren und die Formulierung des Masterplans, der neben einer Aufwandseinschätzung auch fundierte Aussagen zu zwei weiteren Aspekten enthalten wird: Priorisierung und technische Infrastruktur.

Priorisierung

Im Hinblick auf eine Priorisierung stellen sich bislang folgende Fragen:

- *Erschließungsgrad*: Neben sehr gut erschlossenen Beständen gibt es auch Bestände mit älteren, kürzeren oder gar nicht vorhandenen Beschreibungen – welche Rolle soll der Erschließungsgrad bei der Digitalisierung spielen?
- *Bestandsgröße*: Zerstreute Kleinsammlungen, mittlere Sammlungen, größere Sammlungen – welche Rolle spielt die Bestandgröße?
- *Nutzungsbeschränkungen*: Sollen Zimelien und andere besonders wertvolle oder fragile Handschriften mit stark eingeschränkter Nutzbarkeit bevorzugt werden?
- *Inhalt*: Illuminierte Handschriften zu digitalisieren ist besonders naheliegend. Welchen Stellenwert soll die Digitalisierung von Texthandschriften haben, die für die Wissenschaft wichtig, aber für ein breiteres Publikum weniger „attraktiv“ sind?
- *Institutioneller Rahmen*: Soll die Förderbarkeit von Digitalisierungsprojekten von der Nutzbarkeit einer etablierten Digitalisierungsinfrastruktur abhängig sein?
- *Wissenschaftsbezug*: Wie können die Bedürfnisse aktueller Forschungsvorhaben berücksichtigt werden? Wie können die Ergebnisse von Forschungsprojekten in bibliothekarische Ressourcen und Fachdatenbanken eingebunden werden?

Technische Infrastruktur

Zentraler Zugriffspunkt für Digitalate wie für zugehörige Meta-, Erschließungs- und Strukturdaten ist *Manuscripta Mediaevalia* (<http://www.manuscripta-mediaevalia.de>). Das Webportal soll zu einer virtuellen Forschungsumgebung ausgebaut werden.

Ausblick

Von Anfang an sollen Partner aus der Wissenschaft und Informationsinfrastruktur einbezogen werden: Im Oktober 2014 wird eine erste Tagung mit ca. 85 Teilnehmern an der BSB stattfinden, für März 2015 ist eine zweite Tagung zur Evaluierung der Ergebnisse geplant. Im November 2015 wird die Pilotphase abgeschlossen sein und ein tragfähiger Masterplan vorliegen.

Digital Humanities - methodischer Brückenschlag oder "feindliche Übernahme"?

Chancen und Risiken der Begegnung zwischen Geisteswissenschaften und Informatik

Vorschlag für eine Sektion:

Vernetzung von historisch-biographischen Lexika und Fachportalen im Linked (Open) Data Framework

- Praxis, Modelle und Optionen

1. M. Lanzinner (Bonn): Einführung und Moderation
2. M. Jorio (Bern): Das Historische Lexikon der Schweiz: vom nationalen zum europäischen Fachportal
3. M. Schattkowsky (ISGV Dresden): Die Vernetzung der „Sächsischen Biografie“ – Praxis und Ausblick
4. Th. Declerck (DFKI), R. Feigl, Ch. Gruber, E. Wandl-Vogt (Wien): Das WHO's WHO der Habsburgermonarchie im Linked Data Framework
5. B. Ebneth, M. Reinert (München): Die Deutsche Biographie - von der Normdatenvernetzung zum Sparql-Endpoint

Einführung und Moderation

Prof. Dr. Maximilian Lanzinner

(Lehrstuhl für Geschichte der Frühen Neuzeit, Rheinische Friedrich-Wilhelms-Universität Bonn,
Historische Kommission bei der Bayerischen Akademie der Wissenschaften)

**Das Historische Lexikon der Schweiz:
vom nationalen zum europäischen Fachportal**

Dr. Marco Jorio

(Chefredaktor des Historischen Lexikons der Schweiz)

In den letzten Jahren hat sich das *Historische Lexikon der Schweiz (HLS)* in deutscher, französischer und italienischer Sprache als Referenzinformationsmittel zur Geschichte der Schweiz etabliert. Die insgesamt rund 111'000 Artikeln (37'000 je Sprachausgabe) von 2600 Autoren behandeln ca. 26'000 Biographien, 2500 Familien, 5500 Orte und 3000 Themata. Mit dem 13. Band wird im Oktober 2014 die reich illustrierte Buchpublikation abgeschlossen. Seit 1998 sind die fertig bearbeiteten Artikel in der elektronischen Ausgabe des HLS, im sog. e-HLS (www.hls.ch), dem weltweit ersten mehrsprachigen Lexikon im Netz, open access abrufbar. Mit dem Aus- und Aufbau des e-HLS entwickelte sich dieses zum nationalen Portal und steht im Zentrum eines umfassenden Informationsnetzes.

Auf der Basis des HLS wird seit 2010 im Auftrag der Schweizer Regierung das *Neue Historische Lexikon der Schweiz* vorbereitet. Dieses wird nur noch digital publiziert und zusätzlich auch multimediale Informationen, u.a. auch von externen Partnern, anbieten. Wie das „alte“ wird es aber mehrsprachig, wissenschaftlich und dem open-access-Prinzip verpflichtet sein. Zur Zeit werden intern und mit externen Partnern (u.a. Eidgenössische Technische Hochschule, Nationalbibliothek, Schweizerische Akademie der Geistes- und Sozialwissenschaften) Fragen wie Austauschformate, Schnittstellen, Normdaten, Georeferenzierungen, Semantic Web etc. abgeklärt.

HLS und DH in der Schweiz

Das HLS ist seit den 1990er Jahren ein Pionier im Bereich der DH (bevor es diesen Namen überhaupt gab!). Im November 2013 hat die Schweizerische

DHd 2014 - Sektion Historisch-biographische Lexika - 2013-12-27.doc /

6

Vernetzung von historisch-biographischen Lexika und Fachportalen / Lanzinner, Jorio, Schattkowsky, Declerck, Gruber, Feigl, Wandl-Vogt, Ebneth, Reinert / 27.12.2013 14:38:00 / 3085 Wörter / 22405 Zeichen

Akademie der Geistes- und Sozialwissenschaften (SAGW) in einer Tagung „Digital Humanities: Neue Herausforderungen für den Forschungsplatz Schweiz“ eine grundsätzliche Standortbestimmung vorgenommen. Themen waren dabei u. a. DH-Forschungsinfrastrukturen, Data mining in den DH, Webservices, Auswahl geeigneter Software, Data Curation, computerbasierte Forschung in geisteswissenschaftlichen Disziplinen, internationale Kooperationen. Dabei soll dem Neuen HLS künftig eine zentrale Rolle im Rahmen der Schweizer Geschichte und der Geschichtswissenschaft eingeräumt werden.

Bereits jetzt gibt es einen engen Austausch des HLS mit anderen Schweizer DH-Projekten bzw. digitalen Internetressourcen, die z.T. in grossem Umfang auf das HLS verlinken. Das 2013 abgeschlossene Buchprojekt Historische Lexikon des Fürstentums Liechtenstein entstand auf der Basis des HLS und wird jetzt mit Unterstützung des HLS digitalisiert. Im Auftrag der SAGW entwickelten die *Diplomatischen Dokumente der Schweiz* (DoDiS) mit dem HLS den Prototyp des Webservice Metagrid für die Online-Vernetzung von geisteswissenschaftlichen Ressourcen, der jetzt ausgebaut wird. Das *Familiennamenbuch* und das *Ortsnamenlexikon* (Glossarium Helvetiae Historicum) werden bereits seit Jahren als Webservices des HLS betrieben. Die Zusammenarbeit mit andern Anbietern wie dem Schweizerischen Institut für Kunstgeschichte (SIKART), den Schweizerischen Rechtsquellen und der Bibliographie der Schweizergeschichte in der Nationalbibliothek ist eingeleitet oder in Diskussion.

Internationale Kooperation

Das *HLS* steht im Austausch mit zahlreichen biographischen (Online-)Lexika in Europa. Aufgrund einer Kooperationsvereinbarung mit der Bayerischen Staatsbibliothek (BSB), der Österreichischen Akademie der Wissenschaften (ÖAW) und der Historischen Kommission bei der Bayerischen Akademie der Wissenschaften (BAdW) betreibt das HLS seit Juli 2009 das mehrsprachige

Biographie-Portal (www.biographie-portal.eu) mit. Um die weitere Vernetzung zu befördern, hat das *HLS* 2011 begonnen, seine Einträge mit den bereits bestehenden Normdaten der *Gemeinsamen Normdatei* (GND) und des *Virtual International Authority File* (VIAF) zu versehen und beteiligt sich seit Ende 2013 auch direkt an der GND-Redaktion.

Die Vernetzung der „Sächsischen Biografie“ – Praxis und Ausblick

Prof. Dr. Martina Schattkowsky

(Institut für Sächsische Geschichte und Volkskunde, Dresden)

Das Institut für Sächsische Geschichte und Volkskunde (ISGV) stellte bereits im Jahr 2005 das Online-Personenlexikon Sächsische Biografie, die bedeutende sächsische Persönlichkeiten vom 10. Jahrhundert bis zur Gegenwart erfasst, ins Internet. Über dieses Portal sind derzeit über 10.500 Personeneinträge und ca. 1.250 Personenartikel recherchierbar.

Im Beitrag des ISGV sollen die derzeitigen Vernetzungsstrategien sowie einige Vorhaben umrissen werden, sowohl intern (mit anderen Online-Projekten des ISGV, insbesondere des „Digitalen Historischen Ortsverzeichnisses von Sachsen“) als auch extern.

Aktueller Stand

Die Sächsische Biografie beteiligt sich seit geraumer Zeit – zum Beispiel im Rahmen der AG Regionalportale – an der Diskussion über die bessere Vernetzung bzw. Zusammenarbeit mit anderen Online-Personenlexika. Ergebnisse konnten zum Beispiel durch die im Mai 2012 erfolgte Implementierung der Sächsischen Biografie in das europäische „Biographie-Portal“ (<http://www.biographie-portal.eu>) erzielt werden.

Austauschformate

Zum Austausch mit anderen Online-Projekten verwendet die Sächsische Biografie zwei Formate: Zum einen werden – mittels XML-Dateien – Kerndatensätze zu Personen, die in der Sächsischen Biografie verzeichnet sind, an das europäische Biographie-Portal geschickt, die Artikel sind über die dortige

Suchmaske abrufbar. Zum anderen wird durch eine BEACON-Datei die Verlinkung zu anderen Projekten mittels Gemeinsamer Normdatei sichergestellt. Zudem ist die Implementierung von Perma-Links geplant, um die langfristige Abrufbarkeit der Artikel zu gewährleisten.

Normdaten

Eine Verlinkung zu anderen Onlineprojekten wie Bibliotheken oder Biografien erfolgt mithilfe der Gemeinsamen Normdatei (GND). Diese Links werden mithilfe des BEACON-Formats, das von freiwilligen Mitarbeitern der Wikipedia gewartet und weiterentwickelt wird, direkt in den Personenartikeln angezeigt, die Generierung erfolgt über eine Echtzeit-Abfrage (SeeAlso-Service). Problematisch ist dabei der bisweilen zögerliche Umgang einzelner Institutionen mit der Heraus- bzw. Freigabe von GND-IDs.

Datenanalyse bzw. Georeferenzierung

Eine umfangreiche Georeferenzierung der in den Personenartikeln genannten Orte ist derzeit noch in Planung und sollte zukünftig in Zusammenarbeit mit anderen Portalen einheitlich gelöst werden, um eine Vernetzung dieser Daten analog zur Verlinkung von Personen mittels GND zu ermöglichen. Alle 6.000 in Sachsen in den Grenzen von 1990 gelegenen Orte verzeichnet das ebenfalls am ISGV beheimatete „Historische Ortsverzeichnis von Sachsen“. Hierfür existiert bereits eine institutsinterne Verlinkung (ist aber im Netz noch nicht aktiv).

Personale Relationen

Verlinkungen zwischen Personen, die in Artikeln in der Sächsischen Biografie vorkommen, werden derzeit händisch eingepflegt. Durch eine entsprechende Markierung in den Artikeln ist es zurzeit möglich, Verweise auf andere in der Sächsischen Biografie verzeichnete Personen zu generieren und Kerndaten (Namensformen, Lebensdaten, Status des Artikels) der jeweiligen Personen darzustellen. Familienbeziehungen, Stammbäume o. ä. sind derzeit aber noch nicht darstellbar.

Visualisierung

Eine Geo-Visualisierung ist derzeit noch in Planung. Es ist jedoch – eher mittel- und langfristig – bereits daran gedacht worden, komplexe, das heißt mehrere Kategorien umfassende Abfragen in Kartenansichten darzustellen. Hierfür wäre eine interne Verlinkung zum Historischen Ortsverzeichnis von Sachsen denkbar.

Das WHO´s WHO der Habsburgermonarchie im Linked Data Framework
Überlegungen zum Mehrwert einer integrierten Publikation des
Kronprinzenwerks mit dem Österreichischen Biographischen Lexikon
(ÖBL)

**Thierry Declerck (1), Roland Feigl (2), Christine Gruber (2),
Eveline Wandl-Vogt (3)**

- 1) Deutsches Forschungszentrum für Künstliche Intelligenz (DfKI GmbH)
- 2) Institut für Neuzeit- und Zeitgeschichtsforschung (INZ) @ ÖAW
- 3) Austrian Center for Digital Humanities (ACDH) @ ÖAW

Die 1883 vom österreichisch-ungarischen Kronprinzen Rudolf angeregte „österreichisch-ungarische Monarchie in Wort und Bild“, allgemein als „Kronprinzenwerk“ (KPW) bezeichnet, ist eine umfassende landeskundliche Enzyklopädie in deutscher ([1]) und ungarischer Sprache ([2]). Es gab zwei Redaktionen und dementsprechend eine divergierende sprachliche Realisierung. Für den Beitrag wird im Folgenden auf die deutschsprachige Ausgabe fokussiert.

Das KPW versammelte zu Ende des 19. Jahrhunderts die namhaftesten Wissenschaftler, Schriftsteller und Zeichner der Zeit. Es wurde von 432 Mitarbeitern, darunter auch Kronprinz Rudolf selbst, verfasst; die über 4.500 Illustrationen stammen von 264 Künstlern.

In diesem Beitrag wird der Mehrwert durch Vernetzung und (programmgestützte) Informationserfassung aus unterschiedlichen Quellen, Medien und Domänen am Beispiel der Verlinkung von KPW und dem Österreichischen Biographischen Lexikon (ÖBL, [3,4]) über Linked (Open) Data (LOD, [5]) diskutiert.

Aus dem Text des KPW werden Personennamen und damit verbundene Informationen automatisch extrahiert und - beispielhaft für weitere Vernetzungen mit historisch-biographischen Lexika und Fachportalen sowie weitere unstrukturierte textuelle Ressourcen - mit den vorhandenen

strukturierten Daten vom ÖBL verbunden. Dadurch wird am Beispiel der Mitarbeiter des KPW und der thematisierten Personen der Stellenwert des Werkes für die späte Habsburgermonarchie und darüber hinaus (Anfänge der Ethnographie in Österreich) exemplifiziert.

Die Personendaten aus KPW und ÖBL werden in einem RDF / SKOS – Modell repräsentiert, damit sie verlinkt und integriert werden können. Zu diesem Zweck mussten wir erst die Datenbestände des ÖBL in RDF / SKOS portieren, wobei auch das bei der Gemeinsamen Normdatei (GND, [6]) verwendete RDF/XML Modell berücksichtigt wurde. SKOS wird verwendet, um die Verbindung zu Biographiedaten in der Linked Data Cloud differenzierter und flexibler darzustellen, z.B. verwenden wir die SKOS „mapping properties“ *broaderMatch*, *narrowerMatch*, *exactMatch*, *relatedMatch*, *closeMatch*, zusätzlich zu *owl:sameAs*, um Verbindungen zwischen den verschiedenen Datensätzen zu repräsentieren.

Unsere Arbeit zielt darauf, die integrierten Daten aus KPW und ÖBL mit entsprechenden bestehenden RDF-Datensätzen internationaler Netzwerke, z. B. jenem der Deutschen Biographie ([7]), zu verlinken und vergleichen.

Im Zentrum der Betrachtung steht die Diskussion des Mehrwerts durch standardisierte und optimierte Schnittstellen: Am Beispiel konkreter Personen (Karl v. Siegl, der mit zahlreichen Zeichnungen zu dem KPW beigetragen hat, aber der auch von einem anderen Autor beschrieben wird, und der Familie Šubic) diskutieren wir unterschiedliche Repräsentationen der vorliegenden Daten in ausgewählten Modellen (GND [6], DFKI-BiographieOntologie [8], foaf [9], VIAF [10]) und schlagen eine Lösung für eine gemeinsame datenübergreifende Verlinkung und Analyse biographierelevanter Daten vor.

Dies setzt eine genaue Analyse der jeweils ausgewählten Repräsentationsdichte voraus. So zum Beispiel verwendet für die Repräsentation des Todesdatums die GND

„<rdaGr2:dateOfDeath>(JJJJ|JJJJ-MM-TT)</rdaGr2:dateOfDeath>“.

Im ÖBL werden die Angaben für Geburts- und Todesdatum nur als textueller Inhalt eines XML Elements „Kurzdefinition“ angegeben:

„<Kurzdefinition>Šubic Jurij (Georg), Maler, Zeichner und Illustrator.
Geb. Pölland, Krain (Poljane nad Škofjo Loko, SLO), 13. 4. 1855; gest.
Leipzig, Sachsen (D), 8. 9. 1890; röm.-kath.</Kurzdefinition>“

Die DFKI Biography Ontology verwendet dagegen in einer konsistenten Art und Weise nur die xsd Datentypen und setzt „xsd:gYear“ ein, wenn nur das Jahr eines Ereignis bekannt ist, aber nicht Monat und/oder Tag. Ansonsten wird der Datentyp „xsd:date“ verwendet.

Unsere Empfehlung ist es hier, alle datumsrelevanten Informationen auf eine standardisierte Repräsentation abzubilden, so dass wir in der Lage sind, die Inhalte der verschiedenen Quellen zu vergleichen und zu integrieren. Diese Bemerkung gilt nicht nur für Datumsausdrücke, sondern für alle Informationen, die in eine strukturelle Repräsentation überführt werden können, so dass die Datensätze nicht nur über die GND IDs verlinkt werden können, sondern auch umfassend integriert werden können. Wir erwarten von diesem Schritt auch eine genauere Identifizierung von gleichen Personen über Datensätze hinweg.

Über diese Repräsentationsanalysen hinaus, die primär zu Fragen der Publikation der Daten in der Linked Data Cloud gehören, zielen die vorgestellten Arbeiten und Ergebnisse auch darauf ab, aus dem KPW ein bilinguales, paralleles Online-Corpus von Personennamen zu extrahieren und das modifizierte und erweiterte ÖBL als Forschungsinfrastruktur im CLARIN.AT-Netzwerk einzubetten.

Referenzen:

- [1] Die österreichisch-ungarische Monarchie in Wort und Bild, begonnen auf Anregung und unter Mitwirkung von Rudolf von Habsburg, fortgesetzt unter dem Protektorat von Erzherzogin Stephanie. Siehe digitalisierte Edition bei

http://austria-forum.org/af/Web_Books/Kronprinzenwerk (zuletzt eingesehen am: 11.12.2013).

[2] Az Osztrák-Magyar Monarchia írásban és képben.

<http://www.tankonyvtar.hu/hu/tartalom/tkt/osztrak-magyar/adatok.html> (zuletzt eingesehen am: 11.12.2013).

[3] Österreichisches biographisches Lexikon: 1815-1950. Wien / Graz. 1957-.

[4] Österreichisches biographisches Lexikon: 1815-1950 Online-Edition. 2003-.

<http://www.biographien.ac.at/oebi?frames=yes> (zuletzt eingesehen am: 11.12.2013).

[5] Linked (Open) Data: <http://linkeddata.org/> (zuletzt eingesehen am: 11.12.2013).

[6] Gemeinsame Normdatei:

http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html (zuletzt eingesehen am: 11.12.2013).

[7] Neue Deutsche Biographie: <http://www.deutsche-biographie.de> (zuletzt eingesehen am: 11.12.2013).

[8] DFKI Biography Ontology: <http://www.dfki.de/lt/onto/biography.owl> (zuletzt eingesehen am: 11.12.2013).

[9] Friend Of A Friend Modell: <http://www.foaf-project.org/> (zuletzt eingesehen am: 11.12.2013).

[10] Virtual International Authority File: <http://viaf.org/> (zuletzt eingesehen am: 11.12.2013).

Die Deutsche Biographie - von der Normdatenvernetzung zum Sparql-Endpoint

Dr. Bernhard Ebneth, Matthias Reinert

(Historische Kommission bei der Bayerischen Akademie der Wissenschaften)

Der Vortrag umreißt kurz die Bemühungen der Redaktion der Neuen Deutschen Biographie in Zusammenarbeit mit der Bayerischen Staatsbibliothek (BSB), das von ihr bearbeitete biographische Lexikon von der Printfassung her fortzuentwickeln, es um eine Internetpräsentation zu ergänzen mit der Zielrichtung, es auf eine rein digitale Publikation umzustellen.

Meilensteine der Entwicklung, die durch die Kooperation mit der BSB und mit Förderung durch die DFG erreicht wurden, sind

- Kumulierung und Online-Bereitsellung der Register zur NDB und dem Vorgängerlexikon ADB (seit 2001)
- Image-Scans der gedruckten Bände der ADB und NDB (seit 2001 bzw. 2008)
- Volltext-Erfassung und XML-Kodierung der gedruckten Bände (bis 2009)
- Abgleich der Personennamen gegen die GND (damals PND) und Ergänzung derselben (seit 2008)
- Digitalisierung der Arbeitskartei der Redaktion und GND-Abgleich (2010/11)
- Integration zentraler personenbezogener Bestände (seit 2011)
- Aufbau einer RDF-Schnittstelle (seit 2011)
- computerlinguistische Aufbereitung der Artikeltexte und Genealogien

Die Gemeinsame Normdatei (bis 2012 für Personen: PND) der Bibliotheken und -verbünde, koordiniert von der Deutschen Nationalbibliothek, erweist sich als erfolgreiches Erschließungsinstrument für Personen und setzte sich ab 2010 auch in anderen Online-Angeboten und in den Verbünden durch. Mittlerweile kann

die GND-Erschließung auch für Orts- und Sacherschließung mit bibliothekarischem Anschluss genutzt werden und die GND schlägt über eindeutige Identifier eine Brücke in das Semantic Web.

Die retrospektiven Aufbereitungsarbeiten schlagen auf die redaktionelle Arbeit nieder: mittlerweile werden bei den täglichen Auswertung und Neuerfassung in der Redaktion bereits die GND-Einträge zu den Namen/Personen geprüft, die GND-Nummer und ggf. Namensvarianten übernommen.

Mit eigenen IDs insbesondere aber der Hilfe der GND-IDs, die sukzessive teils in Eigenleistung der BSB und der NDB auch in neu erscheinenden Bänden für die Personen des Registers nachgetragen wird, ist die Kumulierung der Einträge aus verschiedenen eigenen Ressourcen für www.deutsche-biographie.de möglich. Zudem lassen sich die personenbezogenen Ressourcen unserer Kooperationspartner (Bundesarchiv, Deutsches Literaturarchiv Marbach, Foto Marburg, Germanisches Nationalmuseum, Deutsches Rundfunkarchiv, Deutsches Museum bis jetzt) leicht integrieren.

Darüber hinaus kann die GND auch grundlegend zur computerlinguistischen Analyse beitragen. Aus ihr lassen sich Wörterbücher für Institutionen, Orte (Gebietskörperschaften) und Namen (Vor- wie Nachnamen) leicht generieren und verarbeiten. Hat man - wie bei uns mit Hilfe lokaler Grammatiken - Named Entities im Text gefunden (NER), kann die GND auch bei einem weiteren Schritt - der Disambiguierung (NED) unterstützend herangezogen werden.

Die beiden Schritte NER und NED führen zu schematisierten Aussagen über die Inhalte der Biographien: Wer war Schüler oder Lehrer der Biographierten? Welche IDs und Verknüpfungen zu anderen Personen im Korpus ergeben sich? Welches sind die Geburts-, Sterbe- und Wirkungsorte? Wie lassen sie sich auf Karten darstellen? (Um das letzte zu erreichen werden die Orte zudem gegen Openstreetmap und Geonames abgeglichen; die GND hat hier zur Zeit noch keine Geokoordinaten und relativ wenige Einträge.)

Mit Hilfe schematisierter Aussagen, die in einem Triple-Store vorgehalten werden, lassen sich nun neue, bspw. netzwerkanalytische Fragestellungen an das Biographien-Korpus richten. Resultate können auch in Visualisierungen dargestellt werden, z. B. in Ego-Netzwerken.

Literatur

- GND: (http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html)
- BEACON Link dump format, Draft 2012
(<http://gbv.github.io/beaconspec/beacon.html>)
- Guenthner, Franz; Maier, Petra (Hgg.): Das CISLEX Wörterbuchsystem. München 1994. (<http://www.cis.uni-muenchen.de/download/cis-berichte/94-076.pdf>).
- Geierhos, Michaela: BiographIE - Klassifikation und Extraktion karrierespezifischer Informationen. München 2010.
- Brümmer, Martin: Realisierung eines RDF-Interfaces für die Neue Deutsche Biographie, <http://skil.informatik.uni-leipzig.de/blog/historie/skil2011/zusammenfassungen-der-beitrage/#vortrag4>
- Ebneth, Bernhard: Aktueller Stand der Genealogien in der Neuen Deutschen Biographie – Arbeit mit der Online-Version, 2012, (<http://www.ndb.badw-muenchen.de/Genealogentag-NDB-2012.pdf>)
- Ebneth, Bernhard: Das europäische Biographie-Portal mit Allgemeiner Deutscher Biographie und Neuer Deutscher Biographie Online, in: Catalogus Professorum Lipsiensis. Konzeption, technische Umsetzung und Anwendungen für Professorenkataloge im Semantic Web, hg. v. Ulf Morgenstern u. Thomas Riechert, Leipzig 2010, S. 159-168.

Digital Humanities - methodischer Brückenschlag oder "feindliche Übernahme"?

Chancen und Risiken der Begegnung zwischen Geisteswissenschaften und Informatik

Vorschlag für eine Sektion:

Vernetzung von historisch-biographischen Lexika und Fachportalen im Linked (Open) Data Framework

- Praxis, Modelle und Optionen

1. M. Lanzinner (Bonn): Einführung und Moderation
2. M. Jorio (Bern): Das Historische Lexikon der Schweiz: vom nationalen zum europäischen Fachportal
3. M. Schattkowsky (ISGV Dresden): Die Vernetzung der „Sächsischen Biografie“ – Praxis und Ausblick
4. Th. Declerck (DFKI), R. Feigl, Ch. Gruber, E. Wandl-Vogt (Wien): Das WHO's WHO der Habsburgermonarchie im Linked Data Framework
5. B. Ebneth, M. Reinert (München): Die Deutsche Biographie - von der Normdatenvernetzung zum Sparql-Endpoint

Einführung und Moderation

Prof. Dr. Maximilian Lanzinner

(Lehrstuhl für Geschichte der Frühen Neuzeit, Rheinische Friedrich-Wilhelms-Universität Bonn,
Historische Kommission bei der Bayerischen Akademie der Wissenschaften)

Vernetzung von historisch-biographischen Lexika und Fachportalen im Linked (Open) Data Framework

- Praxis, Modelle und Optionen

Die intelligente Vernetzung zwischen autonomen Internetressourcen erhöht nicht nur Sichtbarkeit und Reichweite der einzelnen beteiligten Projekte, sondern generiert durch neue Webservices und die Aggregation von Daten auch einen Mehrwert für neue Erkenntnisse und Fragestellungen. Als aktiver Beitrag zur Forschungsunterstützung werden in den Geschichtswissenschaften auf regionaler, nationaler und europäischer Ebene derzeit neue Informationssysteme und Infrastrukturen geplant und entwickelt. Für heterogene und multimediale personenbezogene Ressourcen wie Fach- und Regionalportale, Lexika, Biblio- und Mediographien, Editionen, Regesten sowie Quellen-, Bild-, Ton- und Filmnachweise sowie verschiedenartige Digitalisate soll in dieser Sektion vorgestellt und diskutiert werden, wie eine entsprechende Vernetzung und Integration zwischen diesen Ressourcen möglichst systematisch, effizient und stabil geleistet werden kann. Dabei soll für die beteiligten Partner die Autonomie und Kontrolle über ihren jeweiligen Content in vollem Umfang gewahrt bleiben und für den User klar erkennbar sein, zu welchen Teilen und aus welchen Quellen der Portalinhalt zusammengestellt ist. Aktuelle Beispiele wären Metagrid.ch, die Deutsche Biographie oder das Biografisch Portaal van Nederland.

Für Fragen und Probleme der Vernetzung wie die Integration und kollaborative Forschung an personenbezogenen Daten sollen am Beispiel von historisch-biographischen Lexika bzw. Fachportalen die folgenden Aspekte untersucht und diskutiert werden:

Datenformate: Welche Rolle kommt Linked (Open) Data im Prozess der Erstellung, Rezeption und Publikation zu? Welche Rückwirkungen auf Dokumentenformate und Metadaten ergeben sich daraus?

Austauschformate: Welche Austauschformate sind zu empfehlen? Wie kann eine langfristige Kompatibilität von archivischen, bibliothekarischen und forschungsnahen Standards bei unterschiedlichen Zugängen und Interpretationen gewährleistet werden?

Normdaten: Wie können persistente Authority Files / Normdaten (wie GND, VIAF) für die langfristige Vernetzung und Datenintegration effizient eingesetzt werden? Welche Anwendungserfahrungen und Probleme ergeben sich?

Datenanalyse: Was können Georeferenzierung und computerlinguistische Analyseverfahren leisten?

Personale Relationen: Wie können Beziehungen zwischen Personen untereinander sowie zu Wirkungsorten, Körperschaften (Institutionen, Organisationen) und Werken eindeutig ermittelt und allgemein verwendbar kodiert und ausgetauscht werden?

Visualisierung: Wie können Beziehungen allgemein verständlich visualisiert werden? Welche Aussagekraft haben Visualisierungen heterogener Daten? Wie lassen sich sinnvolle Fragestellungen möglichst breit durch Datenlieferanten bedienen?

Forschungskommunikation: Wie weit sind Social Media-Komponenten für die Vermittlung und Reintegration von neuen Forschungsergebnissen und Fragestellungen hilfreich? Welche Tools, APIs und Services sind im biographisch-lexikalischen Bereich empfehlenswert? Wie können Utility und Usability bei begrenzten finanziellen und personellen Mitteln langfristig verbessert werden.

Rechte und Attribution: Wie lassen sich Daten über institutionelle Grenzen und domänenübergreifend austauschen, gemeinsam nutzen und gemäß den Regeln guter wissenschaftlicher Arbeit angemessen zitieren? Wie werden Datenaufbereiter, -kodierer und -analysten in den geisteswissenschaftlichen Forschungsstrukturen honoriert?

Für die Gründung und Fortführung geschichtswissenschaftlicher und lexikalischer Projekte ist das geänderte mediale Umfeld von Beginn an mitzudenken und sowohl für die Etablierung als auch für die dauerhafte Pflege eine personell und finanziell angemessen ausgestattete IT-Kompetenz einzuplanen. In den neuen (internationalen) Forschungs-, Publikations- und Kommunikationsstrukturen sollte eine langfristige Bereitstellung und Dokumentation der Forschungsdaten und -ergebnisse garantiert sein. Für den internationalen und interdisziplinären Austausch entstehen derzeit geeignete Foren (z. B. Zentren für digitale Geisteswissenschaften, Blogs, Wissenschaftsportale).

Die digitalen Geschichtswissenschaften im deutschsprachigen Raum können und sollten Erfahrungen und Perspektiven in anderen Ländern in Informatik, Editions- und Naturwissenschaften sowie aus benachbarten Disziplinen wie Philologien, Musikwissenschaften, Altertumswissenschaften, Historischen Grundwissenschaften, Theologie und Kunstgeschichte sowie in Archiv- und Bibliothekswesen sowie Dokumentologie und Museologie produktiv berücksichtigen.

Referenzen:

Biographie-Portal <<http://www.biographie-portal.eu>>
(eingesehen: 18.12.2013)

Biografisch Portaal van Nederland <<http://www.biografischportaal.nl/>>
(eingesehen: 17.12.2013)

Metagrid.ch <<http://www.metagrid.ch/>> (eingesehen: 17.12.2013)

Historisches Lexikon der Schweiz <<http://www.hls-dhs-dss.ch>> (eingesehen:
18.12.2013)

Österreichisches Biographisches Lexikon 1815-1950 Online-Edition /
Österreichisches Biographisches Lexikon ab 1815 (2. überarbeitete Auflage –
online) <<http://www.biographien.ac.at>>
(eingesehen: 18.12.2013)

Deutsche Biographie <<http://deutsche-biographie.de>>
(eingesehen: 18.12.2013)

Sächsische Biografie <<http://saebi.isgv.de>>
(eingesehen: 18.12.2013)

Kommentiertes Digitales Korpus Deutscher Rechts- und Gesetzestexte

- D-LEX -

Florian Kuhn

Universität Potsdam

Institut für Linguistik

AB Angewandte Computerlinguistik

fkuhn@uni-potsdam.de

Dankmar W. Enke

Universität Tübingen

Seminar für Sprachwissenschaft

AB Quantitative Linguistik

dankmar.enke@uni-tuebingen.de

Antje Baumann

Bundesministerium der Justiz

Referat Rechtsprüfung, Sprachberatung,

Allgemeines Verwaltungsrecht

baumann-an@bmj.bund.de

&

Humboldt-Universität zu Berlin

Institut für deutsche Sprache und Linguistik

Universität Passau, 1. Jahrestagung der
Digital Humanities im deutschsprachigen Raum (DHd), 25. - 28.03.2014

1 Einführung

Die Arbeit am und mit dem Recht hat seit jeher nicht nur Juristen, sondern auch Akteure anderer Fachbereiche beschäftigt. Allerdings zeigt sich mit Blick auf die vergangenen Jahrzehnte, dass ein Austausch zwischen diesen Disziplinen kaum stattgefunden hat. Dabei nimmt die Rechtstheorie eine Sonderstellung ein; sie vermochte es Ende der sechziger und Anfang der siebziger Jahre des letzten Jahrhunderts, Impulse aus Sprachphilosophie und Soziologie in juristische Fragestellungen einzubeziehen.

Besonders Juristen und Sprachwissenschaftler pflegen seit Beginn der achtziger Jahre interdisziplinären Austausch im Rahmen von Arbeitsgruppen zur Juridischen Linguistik, gemeinsamen Veröffentlichungen, bei der Einbindung von Linguisten in den Gesetzgebungsprozess sowie in gemeinsamen Studiengängen.

Vor dem Hintergrund unseres digitalen Zeitalters zeichnet sich auch in der Rechtstheorie eine neue Rezeption ab, denn das Interesse an medientheoretischen Konzepten wächst. Rechtslinguisten können und sollen dieses Interesse ihrer Juristischen Kollegen aufgreifen. Auf Grundlage computerlinguistischer und informatischer Anwendungen können sie gemeinsam zu einer empirischen und methodologischen Fundierung beitragen. Diese soll in ein kommentiertes digitales Korpus deutscher Rechts- und Gesetzestexte münden, das wir in unserem Vortrag präsentieren.

2 Sprachprüfung im Deutschen Bundestag

Da die Regelungsinhalte deutscher Gesetze und deren Sprache oft unverständlich sind, führte die Kritik daran zu einer Neuorganisation der Sprachprüfung als Teil der Rechtsprüfung. Seit 2009 prüft der Redaktionsstab Rechtssprache beim Bundesministerium der Justiz alle Gesetz- und Verordnungsentwürfe der Bundesministerien auf sprachliche Richtigkeit und Verständlichkeit. Dass Linguisten einen fremden Blick auf entstehendes Recht haben und andere Fragen stellen, ist gewollt – und verlangt von juristischen und anderen Fachleuten eine neue Art des Sprechens über Texte. Die Sprachprüfung von Gesetzentwürfen verstärkt die (zum Schaffen von neuem Recht prinzipiell nötige) Rechtsvergleichung – wozu die Sprachberater aber auch neuartige Arbeitsmittel benötigen: Juristische Datenbanken sind jedoch bisher weder frei zugänglich und nicht für sprachliche Fragen ausgelegt, die Ergebnisse derartiger Anfragen fallen also zurzeit noch unbefriedigend aus. Gebraucht werden Aussagen über den Gebrauch von Wörtern oder Wortverbindungen (Häufigkeit, Kollokationen, Kontext, Textsorte etc.); dieser pragmatische Aspekt der Sprachbetrachtung fehlt weitgehend in Wörterbüchern und derzeitigen Terminologie-Datenbanken. Aus diesem Zusammenhang erwächst das Bedürfnis nach korpuspragmatischen Methoden und deren Anwendung in digitalisierter Form, um in der Zusammenarbeit mit der juristischen Fachebene möglichst gute Argumente einzusetzen. Ziel bleibt hier somit ein Zugänglichmachen der Texte für Nichtjuristen unter Wahrung fachlicher Präzision sowie die allgemeine Wahrung der sprachlichen Wohlgeformtheit unabhängig jedweder Fachtextsorte. Anzumerken ist, dass die Zugänglichkeit immer mit Bezug auf die tatsächliche Adressatengruppe geschieht. Nicht zuletzt soll somit auch der zeitliche Aufwand bei der Gesetzesformulierung optimiert werden.

3 Primärkorpus zu Rechts- und Gesetzestexten

Um das Korpus für eine zukünftige Verwendung verfügbar zu machen, ist die Repräsentation in einem offenen und vielseitigen Format unablässig. So ist etwa die Kompatibilität zu Architekturen wie ANNIS vorgesehen, da hierdurch auch ein möglicher Austausch mit Studien anderer Textsorten gewährleistet wird. Neben dieser Architektur, die vor allem für computerlinguistische Analysen relevant ist, wird auch ein möglicher Export nach offenen XML-Formaten zur Repräsentation juristischer Daten möglich sein. Ein solches Format stellt Metalex dar, das an mehreren europäischen Universitäten entwickelt und gepflegt wird.

Aufbauend auf bereits erfolgten Vorarbeiten (Kuhn 2009, 2010) werden die für eine semantische Kategorisierung von Rechtstermini textlinguistisch relevanten Informationen aus einem Sekundärkorpus mit Gerichtsentscheidungen durch eine automatenbasiertes mehrbeiges Parsingverfahren erschlossen. Termini, die hier in ihrem unmittelbaren Anwendungskontext auftreten, können im Primärkorpus durch Querverweise zu diesen Kontexten bereichert werden.

4 Digital Humanities und Digital Law?

Durch eine Architektur wie ANNIS kann ein Rechts- und Gesetzeskorpus in einem offenen Format digital zugänglich gemacht werden. Hierdurch wird es auch möglich, einen Austausch mit anderen geisteswissenschaftlichen Inhalten zu fördern, da dies durch eine textsortenunabhängige, linguistisch motivierte Repräsentation gegeben ist. Ebenfalls wird durch das offene Format eine von kommerziellen Interessen unabhängige Niederlegung der Daten möglich.

5 Zusammenfassung und Ausblick

Zur Entstehung des beschriebenen Forschungsprojektes bedarf es einerseits des Aufbaus und der Zugänglichmachung geeigneter Rechts- und Gesetzescorpora, die das gesamte Spektrum dogmatischer Textsorten beinhalten (Entscheidungen der Oberinstanzgerichte sowie rechtswissenschaftliche Kommentarliteratur). Andererseits werden etablierte korpuslinguistische Verfahren, Werkzeuge und Algorithmen dahingehend angepasst und weiterentwickelt, dass

sie für die spezialisierte juristische Textpraxis unter verschiedenen institutionellen Rahmenbedingungen angewendet werden können .

Die bereits oben aufgeworfenen interdisziplinären Berührungspunkte lassen sich nur über einen fruchtbaren, offenen Dialog unterschiedlicher Disziplinen, insbesondere zwischen Rechtswissenschaft, Sprachwissenschaft, Computerlinguistik, Informatik und Medienwissenschaft sowie zwischen Wissenschaftlern und Praktikern der jeweiligen Bereiche bearbeiten. An einer solchen Zusammenarbeit besteht reges Interesse. Verbundprojekte, vor allem im angelsächsischen Raum, belegen dieses Bedürfnis und zeigen bereits erste Ergebnisse. Wir freuten uns, im Rahmen der ersten Jahrestagung der Digital Humanities im deutschsprachigen Raum unser Vorhaben vorzustellen und in Diskussionen fruchtbare Verbindungen zu den angrenzenden Fachbereichen herauszuarbeiten.

References

- [1] **Busse**, Dietrich 2000. *Textlinguistik und Rechtswissenschaft*. Berlin: de Gruyter.
- [2] **Carstensen**, Kai-Uwe, Christian Ebert, Cornelia Endriss, Susanne Jekat und Ralf Klابunde (Hrsg.) 2004. *Computerlinguistik und Sprachtechnologie*. Heidelberg: Springer.
- [3] **Engberg**, Jan 1992. Signalfunktion und Kodierungsgrad von sprachlichen Merkmalen in Gerichtsurteilen. *Hermes* 6, 65–82.
- [4] **Hachey**, Ben und Claire Grover 2006. Extractive summarisation of legal texts. *Artificial Intelligence and the Law* 14.
- [5] **Hopcroft**, John E. und Jeffrey Ullman 1979. *Introduction to Automata Theory*. Addison Wesley.
- [6] **Kuhn**, Florian 2010a. A description language for content zones of german court decisions. *Workshop Programme LREC 2010*.
- [7] **Kuhn**, Florian 2010b. A framework for graph-based parsing of German Private Law Decisions. *Business Information Systems Workshops*. Heidelberg: Springer, 292-297.
- [8] **Sipser**, Michael 2006. *Introduction to the Theory of Computation*. Boston: Wadsworth.

Editionspraxis und technologische Entwicklung: Erfahrungen und Schlussfolgerungen aus der elektronischen Edierung von Wittgensteins Nachlass

Christian Erbacher, Universität Bergen (Norwegen)

1. Vom Buch zur CD-ROM

Wittgensteins „spätere Philosophie“ (geschrieben 1929-1951) ist durch seinen über 15.000 Seiten umfassenden Nachlass und daraus hergestellten Editionen vermittelt (für eine vollständige Bibliographie Pichler et al. 2011). In der 50 Jahre spannenden Editionsgeschichte wurden sehr früh Möglichkeiten der Digitalisierung erprobt: Bereits in den 1960er sollten die damals gedruckten Werke elektronisch zugänglich gemacht werden. Ab Mitte der 1970er Jahre wurde intensiv an der computer-gestützten Optimierung der Darstellung in einer gedruckten kritischen Ausgabe gearbeitet. Ab 1990 begann in einem weiteren Großprojekt an der Universität Bergen (Norwegen) die Transkription und Kodierung des gesamten Nachlasses für eine elektronische Edition, die im Jahre 2000 als *Wittgenstein's Nachlass: The Bergen Electronic Edition* (BEE) in Form von 6 CD-ROMs erschienen ist.

Die BEE verdient hinsichtlich der Entwicklung von Digital Humanities besondere Aufmerksamkeit. Sie war nicht nur die erste nahezu vollständige Ausgabe von Wittgensteins Nachlass, sondern das Medium der elektronischen Ausgabe entsprach auch besonders gut den Eigenarten des Corpus, die sich aus Wittgensteins Arbeitsweise ergeben (Pichler und Erbacher 2008). Die BEE kann mit Varianten, Querverweisen und Überschneidungen umgehen und erlaubt durch vielfältige Suchmöglichkeiten nach Kriterien wie Personen, Datum, Formeln u.v.m. eine zielgerichtete Erschließung. Die für Wittgensteins philosophische Schriften wichtige genetische Rekonstruktion wird durch dieses Werkzeug erheblich erleichtert. Gleichzeitig stellte die komplexe Struktur von Wittgensteins Manuskripten eine besondere Herausforderung für die Entwicklung einer software dar, die solche Strukturen handhaben könnte. Für das Projekt der elektronischen Ausgabe von Wittgensteins Nachlass wurde eigens ein Kodierungssystem mit dieser Stärke entwickelt (MECS: Multi-Element-Coding-System, z.B. Huitfeldt 1994). Mit MECS wurde zu technischen Lösungen bei der Kodierung von komplexen Dokumenten beigetragen und bis heute wirkt diese Arbeit zurück in digitale Disziplinen (siehe z.B. <http://mlcd.blackmesatech.com/mlcd/index.html>).

2. Von der CD-ROM ins Internet

Obwohl erst in 2000 erschienen kommt die BEE im rasanten Internetzeitalter in die Jahre. Die Abhängigkeit von Betriebssystemen und Software macht ein Auslaufen der Benutzbarkeit auf aktualisierten Computern absehbar. Früh wurde deshalb am WAB die Umcodierung des Nachlasses in XML, der lingua franca des Internet, ins Auge gefasst. Sie ermöglicht eine länger dauernde Darstellbarkeit und die Kompatibilität mit anderen Inhalten und Technologien des WWW. Die Umcodierung begann während des eContentplus Programms DISCOVERY (Digital Semantic Corpora for Virtual Research in Philosophy). Dabei wurden über 5000 Seiten von Wittgensteins sog. mittlerer Periode in einer normalisierten und diplomatischen Ausgabe mitsamt den entsprechenden Fascimile frei zugänglich ins Internet gebracht (siehe: www.wittgensteinsource.org). Dies kann als ein erster Schritt hin zu einer Internet-Ausgabe von Wittgensteins Nachlass angesehen werden (Pichler 2010).

Durch die Verlagerung des Corpus ins Internet ergeben sich weitere Bearbeitungsmöglichkeiten, die eine mögliche Zukunft des „digital wittgenstein scholarship“ erahnen lassen (vgl. Falch, Erbacher und Pichler, 2013). Einige Beispiele:

1. Verknüpfung mit anderen relevanten Dokumenten, z.B. mit Primärtexten anderer Philosophen und Schriftsteller, Sekundärliteratur oder etwa mit der elektronischen Ausgabe von Wittgensteins *Gesamtbrieftausch*.

2. Komplexeres Durchsuchen und Erschließen des Corpus, z.B. durch regelbasierte linguistische Suche und semantische Suche wie sie für Wittgensteins Nachlass an der Universität München bereits entwickelt worden ist (siehe: <http://wittfind.cis.uni-muenchen.de/>; vgl. den Konferenzbeitrag von Prof. Hadersbeck und Mitarbeitern)
3. Strukturieren und Erschließen des Corpus durch Technologien des „semantic web“, z.B. durch Erstellen einer „Wittgenstein Ontologie“, mit der seit dem Projekt DISCOCERY am WAB experimentiert wird.

Die Verwirklichung dieser Möglichkeiten liegt in naher Zukunft. Im Folgenden soll etwas näher auf die dritte hier erwähnte Ausprägung des digitalen Wittgenstein scholarship eingegangen werden. Dank der Erfahrung im DISCOVERY-Projekt lassen sich die Erwartungen und Schwierigkeiten einer „Wittgenstein Ontologie“ schon etwas konkreter beschreiben.

3. Schlussfolgerungen zur „Ontologisierung“ von Wittgensteins Nachlass aus dem DISCOVERY-Projekt

Eine Ontologie im informationstechnologischen Sinn besteht aus den Elementen eines Bereichs und ihren Beziehungen zueinander. Durch die in ihrer Art definierten Verbindungen zwischen Elementen (Dokumenten) sollen komplexe Suchanfragen möglich werden. Für den Fall von Wittgensteins Nachlass wäre eine entsprechende Fantasie, dass einzelne Textpassagen mit der Information verbunden würden, was sie behandeln oder wie eine Textpassage zu einer anderen philosophisch in Beziehung steht. Man könnte so z.B. den gesamten Korpus von Wittgensteins Nachlass nach Passagen durchsuchen, die „grundlegend für Wittgensteins Behandlung der Frage nach dem Regelfolgen“ sind und erfahren wie diese Passagen zueinander in Beziehung stehen.

Diese Vorstellung genügt, um einige Probleme eines so verstandenen „semantic web“ aufzuzeigen: zunächst würde eine Suche wie die eben beschriebene immer auf Grund der die Suche leitenden Ontologie stattfinden. Suchergebnisse wären also durch die interpretatorischen Urteile verzerrt, die zuvor zu den jeweiligen Zuordnungen in der Ontologie geführt hatten (vgl. Erbacher 2011). Wenn sich die Forscher allerdings bewusst sind, dass die Ontologien die möglichen Ergebnisse ihrer Suchanfragen einschränken und lenken, dann können solche Suchergebnisse sehr interessant sein. Man könnte sich vorstellen, dass man anhand von Wittgenstein-Ontologien führender Kommentatoren Aufschluss über deren Interpretationen bekäme. Die Ontologien würde dann zu einem Austausch über verschiedene Interpretationen mit einem sehr eingeschränkten Vokabular führen, was zur Klarheit über Unterschiede und Konsequenzen beitragen kann. Voraussetzung hierfür ist allerdings, dass die Ontologien selbst sichtbar und zugänglich wären und dass sich die Forscher aktiv mit ihrem Vokabular vertraut machen.

Was das Vokabular zur Festsetzung der Elemente der Ontologie betrifft, so ist hier eine weitere Enttäuschung von voreiligen Erwartungen angebracht: die Rede von den „verstehenden“ oder „denkenden“ Computern aufgrund der „semantischen“ Technologien beruht auf der Möglichkeit, dass die Maschine alle Implikationen der in der Ontologie festgelegten Relationen expliziert. Dies wiederum ist nur möglich, da es sich hierbei um logische Relationen handelt. Nur sind bei der Formalisierung natürlichsprachlicher Verhältnisse zwischen Textpassagen sehr schnell Grenzen erreicht. Das gesamte Potential einer logischen Verknüpfung von Textpassagen kann folglich nur sehr eingeschränkt genutzt werden. Unterscheidet man zudem nicht deutlich genug die Ebenen der natürlichsprachlichen Bezeichnung und logischen Relation in der Ontologie, kommt es schnell zu Verwirrungen und falschen Erwartungen rund um die Begriffe wie „semantic web“ und „verstehende“ und „denkende“ Computer.

Wenn man dagegen weitgehend von der „Interpretation“ der Texte absieht und sich um bescheidenere Relationen bemüht, dann bietet Wittgensteins Nachlass abermals einen aussichtsreichen Testfall für die Verschränkung von komplexer Textstruktur und technologischer

Entwicklung. Eine sich auf historisch-philologische Fakten sich beschränkende Ontologie wäre für die Einsicht in die Genese von Wittgensteins Texten sehr nützlich. Eine einfache Ontologie in diesem Sinne könnte so lauten:

Bereich (Domain): Wittgensteins Nachlass

Elemente (elements): Bemerkungen

Relationen (relations): ist Überarbeitung von

Die technischen und die wissenschaftlichen Voraussetzungen für diese Ontologie sind bereits gegeben.

4. Ein Werkstattbericht aus der Nachfolge des DISCOVERY-Projekts

Die Einsicht zu der gerade beschriebenen Bescheidung ist ein Resultat des Projektes DISCOVERY. Ihre Umsetzung wäre ein großer Fortschritt für die „Ontologisierung“ von Wittgensteins Nachlass. Sie enthielte weniger streitbare Interpretationen und würde der Forschung unmittelbar nutzen. Derweil scheint die fortschreitende technologische Entwicklung stets das Experimentieren mit neuesten Anwendungen zu fordern (vgl. z.B. das Nachfolge-Projekt von DISCOVERY:

http://wab.uib.no/wab_agora.page). Es hat sich bei den hier angesprochenen interdisziplinären Projekten gezeigt, dass die beteiligten Geisteswissenschaftler bereit sein müssen, sich in die Details der verwendeten Technologien einzuarbeiten. Dazu gehört zunehmend auch die Mündigkeit im Umgang mit Daten. Wenn zum Beispiel Dienste von Unternehmen in Anspruch genommen werden, die Daten und Beiträge von Forschern speichern und evtl. zu kommerziellen Zwecken verwenden, dann muss der Forscher bewusst entscheiden, ob er dieser Praxis zustimmen möchte. Neben der Verantwortung zur Information seitens des beteiligten Geisteswissenschaftlers muss allerdings auch die Forderung an die Software-Entwickler gestellt werden, rechtlich fragwürdige Sachverhalte zur Diskussion zu stellen sowie technisch möglichst ausgereifte Produkte zur Erprobung anzubieten. Vor allem aber werden Forscher bereitgestellte Technologien nur dann verwenden und an ihrer Entwicklung Interesse haben, wenn sie für ihre Arbeit von Nutzen sind. Hier aber ist wieder der Geisteswissenschaftler gefragt, den software-Entwicklern zu beschreiben, was von Nutzen wäre. Eine Beschreibung der geisteswissenschaftlichen Praxis ist somit, noch vor jeder technologischen Anwendung, der erste Beitrag der Geisteswissenschaften zur Entwicklung von Digital Humanities. Sich ihrer eigenen Praxis mehr bewusst zu werden ist ein echtes Desiderat geisteswissenschaftlicher Forschung.

Genannte Editionen:

Ludwig Wittgenstein Gesamtbrieftausgabe - Innsbrucker elektronische Ausgabe, edited by M. Seekircher, B. McGuinness and A. Unterkircher for the Forschungsinstitut Brenner-Archiv, Intelex 2004

Wittgensteins Nachlass - the Bergen electronic edition, edited by The Wittgenstein Archives at the University of Bergen, Oxford: Oxford University Press 2000

Weitere Literatur:

Erbacher, C. 2011, Unser Denken bleibt gefragt: Web 3.0 und Wittgensteins Nachlass, in: S. Windholz und W. Feigl (Hrsg.): *Wissenschaftstheorie, Sprachkritik und Wittgenstein*, Heusenstamm: Ontos, 135-146

- Falch, Erbacher und Pichler 2013, Some observations on developments towards the semematic Web for Wittgenstein Scholarship, in: D. Moyal-Sharrock, A. Coliva und V. A. Munz (Hrsg.): *Mind, Language, Action* - Beiträge zum 36. Internationalen Wittgenstein Symposium
- Huitfeldt, C. 1994, Toward a Machine-Readable Version of Wittgenstein's Nachlass: Some Editorial Problems. *Editio : Internationales Jahrbuch für Editionswissenschaft*, **6**, 37-43
- Pichler, A. 2010, Towards the New *Bergen Electronic Edition*, in: N. Venturinha (Hg.): *Wittgenstein After His Nachlass*, Hounds Mills: Palgrave Macmillan, 157-172
- Pichler, A. und Erbacher, C. 2008: „Das „Wittgenstein MS 101 from September 1914-Projekt und das Wittgenstein Archiv an der Universität Bergen“ in: J. Bremer/J. Rothaupt (eds.): *Ludwig Wittgenstein: „przydzielony do Krakowa“/„Krakau zugeteilt“*, Krakau: Ignatianum, 199-242
- Pichler, A., M.A.R. Biggs and Pichler, Biggs and Szeltner 2011: *Bibliographie der deutsch- und englischsprachigen Wittgenstein-Ausgaben*, Wittgenstein-Studien, **2**, 249-286

Dr. Thomas Ernst (Universität Duisburg-Essen)

Jenseits des wissenschaftlichen Werks und des geistigen Eigentums? Die digitale Verbreitung wissenschaftlichen Wissens

Ein Exposé für die 1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd) zum Thema *Digital Humanities - methodischer Brückenschlag oder ‚feindliche Übernahme? Chancen und Risiken der Begegnung zwischen Geisteswissenschaften und Informatik*, vom 25. - 28.03.2014 an der Universität Passau

In den meisten geisteswissenschaftlichen Fächern werden Erkenntnisse vorrangig aus der Analyse sprachlicher Schriften gewonnen, die im Regelfall dem Urheberrecht und seinen Regelungen unterworfen sind. Zur Verbreitung geisteswissenschaftlicher Erkenntnisse haben sich in der ‚Gutenberg-Galaxis‘ spezifische mediale Dispositive etabliert, die die Bewertung und Distribution der Erkenntnisse innerhalb des wissenschaftlichen Spezialdiskurses ermöglichen und kanalisieren, dazu zählen u.a. die Veröffentlichungsformen von Wissenschaftsverlagen, die Hierarchisierung wissenschaftlicher Magazine oder die Abläufe von Peer-Review-Verfahren.

Digitale Medien sowie insbesondere die Potenziale der digitalen Kopie und der Sozialen Medien ermöglichen jedoch andere Formen der Verbreitung wissenschaftlichen Wissens, die tendenziell schneller, kürzer, interaktiver, offener sind. Die Nutzung dieser Potenziale ist in den Geisteswissenschaften allerdings umstritten und umkämpft, wie beispielsweise der *Heidelberger Appell für Publikationsfreiheit und die Wahrung der Urheberrechte* (2009) gezeigt hat, und treffen bis heute auf die vehemente Verteidigungshaltung der meisten Wissenschaftsverlage und -buchhändler, wie jüngst noch die Debatte im Anschluss an die die Erklärung *Open Access: Zeit für einen Neubeginn* vom 19. November 2013 aus dem Börsenverein des Deutschen Buchhandels gezeigt hat.

Vor diesen Hintergründen will der Vortrag die Potenziale und Probleme digitaler Medien bei der Verbreitung wissenschaftlichen Wissens aus einer medienliteraturwissenschaftlichen und diskursanalytischen Perspektive auf einer Metaebene reflektieren. Dabei wird er notwendige Differenzierungen und Problematisierungen der Kategorien ‚Autorschaft‘, ‚Werk‘ und ‚geistiges Eigentum‘ vorstellen, die sich in verschiedenen Forschungsarbeiten im Bereich ‚Literatur und Medienpraxis‘ an der Universität Duisburg-Essen als nützlich erwiesen haben. Der Vortrag geht davon aus, dass sich in den wuchernden Diskursen um das digitale Publizieren sehr unterschiedliche Vorstellungen von wissenschaftlichen, journalistischen und literarischen Werken und Autorschaften niederschlagen, die nicht miteinander vermengt werden sollten.

Wissenschaftliche Schriften sind einem spezifischen Regelsystem unterworfen, das als Spezialdiskurs die Produktion neuer Erkenntnisse gewährleisten soll (Jürgen Link 1997). Dieses Regelsystem produziert verschiedene Widersprüche und Probleme: Es benötigt eine feine Balance zwischen präzisen Verweisen auf bekanntes Wissen sowie Momente ‚originärer‘ und ‚neuer‘ Erkenntnisse. Diese Erkenntnisse werden im Regelfall auf eine individuelle Autorschaft zurückgeführt – der Wissenschaftlername und seine Position im akademischen System verbürgen die Originalität des Gedankens, die als ‚geistiges Eigentum‘ dem Wissenschaftler zugeordnet wird (und von Verlagen publiziert und ökonomisiert wird). Dabei lässt sich allerdings auch in wissenschaftlichen Schriften die „Ego-Pluralität“ (Foucault 1969) von Autorschaft analysieren, die beispielsweise das eigentlich vorherrschende ‚Ich-Tabu‘ in wissenschaftlichen Schriften in Vorworten oder Danksagungen aushebelt, weshalb sich unterschiedliche Manifestationsgrade einer wissenschaftlichen Autorschaft in einem Text differenzieren lassen (Steiner 2009). Wissenschaftliche Erkenntnisse werden im Regelfall – da sie den Anspruch erheben, einen dauerhaften Erkenntniswert zu besitzen – als ‚Werk‘ veröffentlicht, bei dem es sich – so das *Reallexikon der deutschen Literaturwissenschaft* – um ein „fertige[s] und abgeschlossene[s] Ergebnis“ handelt, „das einem Autor zugehört und in fixierter, die Zeit überdauernder Form vorliegt, so daß es dem Zugriff des Produzenten ebenso entzogen ist wie dem Verbrauch durch den Rezipienten.“ (Thomé 2003)

Die digitale wissenschaftliche Kommunikation kann nun in fünf Ebenen unterteilt werden: Erstens können wissenschaftliche Monografien und Aufsätze in digitaler Form verfügbar gemacht werden (u.a. in Datenbanken, auf Bibliotheks- oder Institutsseiten); zweitens können Wissenschaftler/innen ihre Texte – wie beispielsweise Rezensionen – auf digitalen Plattformen frei zur Verfügung stellen (u.a. IASLonline, literaturkritik.de); drittens können sie (teilweise anonymisiert) konkrete Forschungsfragen auf entsprechenden Online-Plattformen kollaborativ bearbeiten (u.a. Guttiplag-Wiki); viertens können sie auf Wissenschaftsblogs (gekürzte oder trivialisierte) Beiträge zu ihren Forschungsergebnissen oder aktuelle Informationen bereitstellen und ggf. mit einer interessierten Öffentlichkeit diskutieren (u.a. dhd-blog.org); fünftens können Geisteswissenschaftler/innen vernetzt und meinungsreich in sozialen Medien kommunizieren (u.a. Twitter, Facebook, academia.edu).

Diese verschiedenen digitalen Formen des öffentlichen Publizierens als wissenschaftliche/r Autor/in lassen sich in sehr unterschiedlicher Weise auf die traditionellen Formen wissenschaftlicher Autorschaft und Werkbegriffe beziehen. Wenn man diese Beziehungsformen skaliert, wäre der o.g. erste Typus noch sehr nah an traditionellen wissenschaftlichen Formaten (er macht sie allerdings freier verfügbar und kollidiert dabei mit den einschränkenden Regelungen des Urheberrechts), während die Ebenen drei bis fünf mit ihren verschiedenen konstitutiven Elementen (z.B. stark kollaborative oder sogar anonymisierte Autorschaft; interaktive Textproduktion; temporärer oder versionierten Charakter eines Textes etc.) sowohl die bisherigen Vorstellungen von ‚wissenschaftlicher Autorschaft‘ als auch von ‚wissenschaftlichem Publizieren‘ erweitert. Es ließe sich an verschiedenen Beispielen zeigen, dass im digitalen Wandel somit die traditionelle Differenz zwischen ‚wissenschaftlichen‘ und ‚populärwissenschaftlichen Autoren‘ (Parr 2008) aufgelöst wird, wobei dieser Schritt ambivalent ist: Einerseits ermöglichen diese neuen Formen digitaler Wissenschaftskommunikation eine größere Öffnung der Wissenschaft zu nicht-akademischen Diskursen (und der Spezialdiskurs ‚Wissenschaft‘ wird teilweise Teil des Interdiskurses ‚Soziale Medien‘), die der geisteswissenschaftlichen Forschung eine neue Form gesellschaftlicher Legitimation ermöglicht. Andererseits stellen ihre Kritiker (u.a. Reuß 2012) die These auf: Je intensiver ein Autor die Potenziale sozialer Medien wie Twitter oder Weblogs mit Kommentarfunktion – ihre Schnelligkeit, Kürze und Interaktivität – nutzt, desto weniger handele es sich überhaupt um eine Form ‚wissenschaftlicher Autorschaft‘.

In einer kurSORischen Analyse literaturwissenschaftlicher Online-Rezensionsplattformen, Weblogs und Tweets sollen zuletzt zwei Thesen belegt werden: Erstens stehen die Potenziale des digitalen wissenschaftlichen Publizierens in einem Konfliktverhältnis zu bestehenden Vorstellungen von ‚wissenschaftlicher Autorschaft‘ und ‚wissenschaftlichem Werk‘, was dazu führt, dass diese Potenziale eines direkteren, interaktiven, offeneren, versionierten Veröffentlichens – selbst auf literaturwissenschaftlichen Online-Angeboten – häufig nur erstaunlich eingeschränkt genutzt werden. Die Auseinandersetzung mit den Potenzialen des digitalen Publizierens in der Wissenschaft benötigt also eine Reflexion der Historizität und des Wandels des wissenschaftlichen Autor- und Werkbegriffs. Zweitens ist der Status des digitalen wissenschaftlichen Publizierens auf kollaborativen Plattformen, in Weblogs und Sozialen Medien insofern prekär, als hier im Regelfall die Grenzen des Spezialdiskurses Wissenschaft überschritten werden. Zwar ist es für Geisteswissenschaftler/innen, die offensiv die digitalen Medien nutzen, konstitutiv, diesen Schritt zu gehen (und zudem auch ihre Forschungsergebnisse frei verfügbar zu machen, trotz der bestehenden Einschränkungen durch das Urheberrecht), allerdings steht die erhöhte methodologische Komplexität geisteswissenschaftlicher Forschung in den Digital Humanities einerseits ihrer erforderlichen Popularisierung im Dialog mit der Öffentlichkeit andererseits aporetisch entgegen. Die Digital Humanities kommen daher zwangsläufig nicht umhin, ihre spezifischen Ansätze, Fragestellungen und Methoden auch selbstreflexiv und wissenschaftshistorisch zu reflektieren und sich – im Bewusstsein der beschriebenen Aporien – auf wissenschaftspolitische Positionen zu verstndigen.

Kontakt

Dr. Thomas Ernst
Universität Duisburg-Essen
Fakultät für Geisteswissenschaften
Germanistik/Literatur und Medienwissenschaft
Universitätsstr. 12
45141 Essen
thomas.ernst@uni-due.de
Telefon: 0201-183-2291

Zur Person

Dr. Thomas Ernst (*1974) studierte in Duisburg, Berlin, Bochum und Leuven/Belgien, war 2005 Gastwissenschaftler der Columbia University of New York, wurde 2008 promoviert von der Universität Trier und arbeitete anschließend als Postdoktorand an der Université du Luxembourg. Seit 2010 ist er wissenschaftlicher Mitarbeiter an der Universität Duisburg-Essen, u.a. im MA-Studiengang ‚Literatur und Medienpraxis‘, dort habilitiert er über die Geschichte des geistigen Eigentums und forscht über die Potenziale und Probleme des digitalen Publizierens. Derzeit ist er zudem Sprecher der ‚AG Potenziale digitaler Medien in der Wissenschaft‘ in der Global Young Faculty III (2013-2015), er initiierte das Weblog *Digitur – Literatur in digitalen Medien* (blogs.uni-due.de/digitur) und organisierte den Workshop *Nach dem geistigen Eigentum? Digitale Literatur, die Literaturwissenschaft und das Immaterialgüterrecht* (10. Januar 2014; www.uni-due.de/ndge).

Uni-Webseite: www.uni-due.de/germanistik/ernst
Twitter: @DrThomasErnst

,Publizieren in digitalen Medien‘ – Veröffentlichungen und Vorträge 2013/2014 (Auswahl)

Bloggen.

Aufsatz in: Matthias Bickenbach/Heiko Christians/Nikolaus Wegmann (Hg.): Historisches Wörterbuch des Mediengebrauchs. Wien; Köln; Weimar: Böhlau, 2014 (zur Veröffentlichung angenommen, im Erscheinen).

Collaborative Authorship.

Kurvvortrag und Talk mit Prof. Dr. Martha Woodmansee und Dr. Jeanette Hofmann. Auf der Konferenz: Literatur digital/digital literature; organisiert von der Humboldt Law Clinic Internetrecht der Humboldt-Universität zu Berlin, dem Fiktion e.V. und dem Haus der Kulturen der Welt; 21.03.2014, Berlin, Haus der Kulturen der Welt (in Vorbereitung).

Geschäftsmodelle der digitalen Literatur: Das Beispiel Crowdfunding und Crowdsourcing und seine Potenziale und Probleme.

Auf der Konferenz: Managing Popular Culture? Zur Entstehung des Populären zwischen Emergenz und Strategie. 6. Jahrestagung der AG Populärkultur und Medien in der Gesellschaft für Medienwissenschaft; 30.1.-1.2.2014, Karlsruhe, Karlshochschule International University (in Vorbereitung).

Nach dem geistigen Eigentum? Die Literaturwissenschaft und das Immaterialgüterrecht.

Vortrag beim Workshop: Nach dem geistigen Eigentum? Digitale Literatur, die Literaturwissenschaft und das Immaterialgüterrecht. Finanziert aus Mitteln des Rektorats der Universität Duisburg-Essen, in Kooperation mit dem MA-Studiengang ‚Literatur und Medienpraxis‘ an der Universität Duisburg-Essen, dem DFG-Graduiertenkolleg 1787 ‚Literatur und Literaturvermittlung im Zeitalter der Digitalisierung‘ an der Universität Göttingen und der AG ‚Potenziale digitaler Medien in der Wissenschaft‘ der Global Young Faculty III; 10.1.2014, Universität Duisburg-Essen (in Vorbereitung).

Jenseits von Experten und Laien? Literaturkritik als ‚User Generated Content‘ – Probleme und Potenziale für Medien, Verlage, Wissenschaft und Schule.

Vortrag auf dem Deutschen Germanistentag 2013 zum Thema: Germanistik für das 21. Jahrhundert. Positionierungen des Faches in Forschung, Studium, Schule und Gesellschaft; 24.9.2013, Christian-Albrechts-Universität Kiel.

Das ‚Werk‘ und seine ‚Versionen‘. Zum (un)abgeschlossenen Status des Texts aus Sicht der Literaturwissenschaft.

Vortrag auf dem Symposium: Eine neue Version ist verfügbar; 11.5.2013, Evgl. Akademie Tutzing. → Podcast (Video): <http://vimeo.com/66025708> (53:11 Min.)

Daten-Adaptation als Analysemethode für geisteswissenschaftliche Forschung. Ergebnisse einer Untersuchung zur wissenschaftlichen Kommunikation in der modernen Physik*

Martin Fechner

Max-Planck-Institut für Wissenschaftsgeschichte und Berlin-Brandenburgische Akademie der Wissenschaften

I. EINLEITUNG

Die aktuellen Entwicklungen auf dem Feld der digitalen Geisteswissenschaften unterstützen die Forschung durch Digitalisierung [1], Benutzung neuer Werkzeuge [2,3], Datenanalyse [4] oder Visualisierungen [5].

Die vielfältigen Möglichkeiten, die sich einerseits aus einer computergestützten Datenhaltung ergeben, werden einem durch die datenintensive naturwissenschaftliche Forschung, etwa im Bereich der Klimawissenschaften [6], vor Augen geführt. Andererseits werden im Internet die Möglichkeiten einer computerunterstützten Datenpräsentation auf professionell gestalteten Webseiten sichtbar, dort wird etwa im Bereich des Datenjournalismus [7,8] zusätzlich eine Interaktion zwischen Rezipient und Datenpräsentation möglich gemacht.

Die Geisteswissenschaften stehen vor der Frage, welchen Nutzen sie aus den vorhandenen Techniken ziehen, ohne dabei den Anspruch aufzugeben, die Forschungsrichtung zu bestimmen, und sich in ihren Forschungsfragen auf solche reduzieren zu lassen, welche die Technik scheinbar vorgibt. Statt also digitale Werkzeuge mit einer bloßen Digitalisierung oder einem automatisch erzeugten Datenkorpus zu verbinden, wird hier ein alternativer Weg vorgeschlagen, der die geisteswissenschaftliche Forschungsfrage an den Anfang stellt, davon die Auswahl und Generierung von passenden Daten abhängig macht und so eine sinnvolle Analyse ermöglicht.

II. PROBLEM DER FORSCHUNGSDATEN

Alle Formen von Daten beschreiben die Forschungsobjekte aus einer speziellen Perspektive. Bei der Auswertung der Daten muss die Forschungsfrage an die Perspektive der Daten angepasst werden, um eine sinnvolle Auswertung zu ermöglichen. Übernimmt die Forschung unreflektiert die vorhandenen Techniken, so macht sie sich letztlich von den vorgegebenen, begrenzten Auswertungsmöglichkeiten abhängig.

So werden in letzter Zeit vielfach Netzwerke erforscht, wohl auch da hier scheinbar einfache Visualisierungsmöglichkeiten bestehen, die beim Auslesen eines entsprechenden Datenkorpus direkt umgesetzt werden können. Vergleicht man aber einige Netzwerkvisualisierungen [5,9], wird ersichtlich, wie komplex die Darstellung tatsächlich ist und wie individuell der Prozess der Visualisierung begriffen werden muss.

Unter dem Stichwort „Big Data“ [10] werden wiederum Möglichkeiten gesucht, große, oft automatisch erzeugte Datenmengen sinnvoll für die Forschung auszu-

werten. Bislang hat sich aber in vielen Fällen noch kein geeigneter Weg gefunden, auf dem sich dieses Ziel sicher erreichen ließe. Als schwierig hat sich zum einen die Homogenisierung heterogener Datenmengen erwiesen, viel problematischer ist allerdings die Beschränktheit der Daten selbst. Es lassen sich nicht beliebige Forschungsfragen an jedem Datenbestand klären.

III. ANSATZ DIESER ARBEIT

Im folgenden soll mit der Daten-Adaptation (engl. ‘data adaptation’) eine neue Methode vorgestellt werden, die vom Autor im Rahmen des wissenschaftshistorischen Dissertationsprojektes „Kommunikation von Wissenschaft in der Neuzeit“ [11] entwickelt wurde. Mit dieser Methode wird durch ein systematisches Vorgehen eine Datensammlung generiert, die einerseits im Einklang mit der Forschungsfrage steht und andererseits durch die Heterogenität und Wiederverwendbarkeit eine Vielzahl von Analysen möglich macht. Daten-Adaptation meint hier keine technische Schnittstelle, um die Austauschbarkeit zwischen verschiedenen Datenbeständen oder technischen Geräten herzustellen, sondern eine konzeptuelle Anpassung des Datenmodells an die eigene Forschungsfrage und die damit verbundene Datengenerierung und -analyse, die einen transparenten Umgang mit Forschungsdaten vorsieht und bisher unerforschbares und damit unentdecktes offenbaren kann.

Die folgende Demonstration basiert auf zwei Fallbeispielen, die vom Autor im Rahmen des genannten Dissertationsprojektes untersucht werden.

IV. METHODE

In der hier vorgestellten Methode wird die Forschung durch eine Anpassung der Daten in mehreren Schritten begleitet. Zunächst werden (1) die Forschungsfragen den Forschungsgegenständen gegenübergestellt und es werden (2) durch die Einordnung in Beschreibungsmodelle die Qualitäten herausgearbeitet, die für die Beschreibung des Forschungsinteresses hilfreich sind. Mit der Entwicklung (3) eines darauf aufbauenden Datenmodells wird dann die Perspektive auf die Forschungsobjekte von der Forschungsfrage vorgegeben und (4) daran angepasste Daten können gesammelt werden. Die über die Objekte verfügbaren Informationen werden dem Datenmodell entsprechend in den Datenkorpus eingefügt. Die so bei der Quellenbetrachtung erhobenen Daten werden erst in der Forschung selbst generiert und stellen keine bloße Digitalisierung oder Kommentierung der Quellen dar. Diese

* Eine detailliertere Beschreibung findet sich in der noch unveröffentlichten Dissertationsschrift des Autors [11].

Adaptation und Anpassung der Daten an die eigentliche Forschungsfrage kann einen reichhaltigen und vielschichtigen Datenkorpus erzeugen. Die anfangs formulierten Forschungsfragen lassen sich dann (5) schließlich explorativ oder (6) auch in Detailbetrachtungen erforschen.

A. Beispiel-Modellentwicklung

Im Rahmen der wissenschaftshistorischen Untersuchung werden Unterschiede und Gemeinsamkeiten der wissenschaftlichen Kommunikation im 19. und 20. Jahrhundert untersucht. Dabei werden für die Fallbeispiele der Entwicklung der Spektralanalyse und des Lasers die Fragen gestellt, wie aus Forschung im Labor anerkannte Entdeckungen wurden und welche Rolle die verschiedenen Medien spielten.

Aufbauend auf verschiedene Öffentlichkeitsmodelle konnte ein eigener Ansatz erarbeitet werden, in dem von Kommunikationsräumen gesprochen wird, die mit verschiedenen Eigenschaften ausgestattet sind. Ein Kommunikationsraum lässt sich hiernach etwa durch die Qualitäten Größe, Symmetrie, Organisationsgrad und thematische Ausrichtung beschreiben.

B. Datenmodell

Anschließend an den theoretischen Forschungrahmen wurde ein Datenmodell entwickelt, mit welchem die zu untersuchenden Publikationsquellen beschrieben werden können. Das Datenmodell kombiniert dabei mehrere Anforderungen. Zunächst kann es die allgemeinen Angaben zu Autor, Titel, Jahr etc. als leicht verfügbare Basisbeschreibung festhalten. Daneben wurden eigene Kategorien gebildet, mit denen die Quellenobjekte dem Forschungsmodell zugeordnet werden können. Schließlich ist es mit dem Datenmodell auch möglich, Detailbeschreibungen ohne technischen Aufwand mitzunotieren. Da das Datenmodell einem Datenobjekt keine feste Anzahl von Eigenschaften zuordnet, wird mit einem XML-Schema [12] gearbeitet, das alle obigen Anforderungen erfüllt.

C. Datenerhebung und Quellenauswahl

Die Quellenauswahl wurde im Einklang mit dem Forschungsmodell vorgenommen und es wurden jeweils für die Fallbeispiele der Spektralanalyse und des Lasers die Buchpublikationen und wissenschaftlichen Zeitschriften der ersten zehn Jahre seit der entsprechenden Entdeckung untersucht. Bei der Datenerhebung wurde auf die bestehenden Daten aus verschiedenen Bibliothekskatalogen zurückgegriffen und eigene Informationen wurden ergänzt.

D. Analysevorgehen

An den erhobenen Datenbestand können durch Einsatz von XML-Techniken, wie etwa XSL-Transformationen, Reguläre Ausdrücke und Xpath, verschiedene Abfragen

gestellt werden, die die Resultate direkt oder unter Benutzung von Analyseprogrammen als Textausdrücke, Tabellen oder Graphen präsentieren. Natürlich sind nur Abfragen sinnvoll, die im Einklang mit der formulierten Forschungsperspektive stehen.

E. Ergebnisse

In den untersuchten Fallbeispielen werden durch die große Menge von über 1.500 untersuchten Publikationen viele Auswertungen der Daten möglich. So wird durch eine statistische Betrachtung die Unterscheidung der verschiedenen Medien Buch und Zeitschriftenartikel unterstrichen und zusätzlich wird auf eine Unterscheidung zwischen langen und kurzen Büchern hingewiesen. Die Auswertung zeigt auch eine medienabhängige Ausbreitung der Themen, indem sie Spezialisierungs- und Populärisierungsprozesse aufdeckt.

Weiterhin geben die Daten Aufschluß über geographische Verbindungen in der Verlagslandschaft der wissenschaftlichen Literatur und man kann neben einer festen Struktur auch die Veränderungen zwischen den Jahrhunderten nachweisen. Schließlich können die Daten implizite Verbindungen zwischen den Publikationen aufdecken und Hinweise auf Verhaltensweisen von Wissenschaftlern geben.

F. Probleme der Daten-Adaptation

Mögliche Probleme bei der Methode der Daten-Adaptation sind schlecht gebildete Kategorisierungen, die keine eindeutige Zuordnung zulassen, oder eine unzureichende Quellenbasis.

Ob das erste Problem auftritt und das gewählte Datenmodell nicht vollständig mit den Quellen vereinbar ist, kann schon bei der Erhebung der Daten schnell festgestellt werden. Empfehlenswert ist, das Modell an möglichst klar zu beschreibende Qualitäten der Quellen anzupassen. Bei einer möglichen und nicht aufzulösenden Uneindeutigkeit kann das Modell entweder in diesem Punkt eine offenere Beschreibung vorsehen oder es kann wie in sozialwissenschaftlichen Feldanalysen mit einem Reliabilitätsfaktor gearbeitet werden. Dem zweiten Problem kann nur durch eine umfangreiche Recherche oder einer gezielten Verengung des Untersuchungsgegenstand begegnet werden.

V. ZUSAMMENFASSUNG UND AUSBLICK

Die im Rahmen der Arbeit erhaltenen Ergebnisse demonstrieren, wie sich mit dem Ansatz der Daten-Adaptation neue Erkenntnisse durch Datanalyse auch im Rahmen einer hermeneutischen und geisteswissenschaftlichen Arbeit gewinnen lassen. Durch ein mehrstufiges Verfahren, welches von fachlichen Überlegungen geleitet wird, können Forschungsfragen bearbeitet werden und implizite Strukturen offenbart werden.

Hier konnte gezeigt werden, wie sich dieses systematische Verfahren im Rahmen einer wissenschaftshistorischen Untersuchung anwenden lässt und wie sich neue Erkenntnisse und Zusammenhänge aufdecken lassen. Weitere Forschungen müssen zeigen, welche Erkenntnisse sich mit dieser Methode auch in anderen Forschungszweigen gewinnen lassen können. Eine genaue Ausformulierung der Methode folgt in der Dissertationsschrift.

VI. LITERATUR*

- [1] Digitalisierung und Digitale Sammlungen, Münchner Digitalisierungszentrum (MDZ) an der Bayerischen Staatsbibliothek. – <<http://www.muenchener-digitalisierungszentrum.de>>
- [2] Dariah-DE Tools für die Geistes- und Kulturwissenschaften. – <<https://de.dariah.eu/tools>>
- [3] Werkzeuge und Dienste, CLARIN-D. – <<http://de.clarin.eu/de/sprachressourcen/werkzeuge-und-dienste.htm>>
- [4] Stéfan Sinclair und Geoffrey Rockwell, *Teaching Computer-Assisted Text Analysis: Approaches to Learning New Methodologies*, in: Digital Humanities Pedagogy: Practices, Principles and Politics, hg. v. B. D. Hirsch, Cambridge 2012. – <<http://dx.doi.org/10.11647/OBP.0024>>
- [5] Marian Dörk, Heidi Lam und Omar Benjelloun, *Accentuating Visualization Parameters to Guide Exploration*, in: CHI 2013: Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems, ACM, May 2013, S. 1755-1760. – <<http://mariandoerk.de/accentuation/>>
- [6] Klimawirkung und Vulnerabilität – Forschungsbereich II, Potsdam-Institut für Klimafolgenforschung. – <<http://www.pik-potsdam.de/forschung/klimawirkung-vulnerabilitat>>
- [7] Lorentz Mazat, Datenjournalismus, veröffentl. bei bpb.de am 26. Oktober 2011. – <<http://www.bpb.de/gesellschaft/medien/opendata/64069/datenjournalismus>>
- [8] Stefan Plöchinger, Datenjournalismus und digitale Infografiken, veröffentl. im SZblog am 8. März 2013. – <<http://www.sueddeutsche.de/kolumne/datenjournalismus-und-digitale-infografiken-entdecken-sie-unseren-datagraph-1.1619138>>, <<http://sz.de/datagraph>>
- [9] Ulrik Brandes, Linton C. Freeman und Dorothea Wagner, *Social Networks*, in: Handbook of graph drawing and visualization, hg. v. R. Tamassia, London 2010. – <<http://nbn-resolving.de/urn:nbn:de:bsz:352-244311>>
- [10] „Big Data for the Humanities“ Workshop im Oktober 2013, während der 2013 IEEE International Conference on Big Data. – <<http://bighumanities.net/events/ieee-bigdata-2013/workshop-program>>
- [11] Martin Fechner, Diss., Kommunikation von Wissenschaft in der Neuzeit, in Arbeit.
- [12] Tim Bray, Extensible Markup Language (XML) 1.0 (Fourth Edition) – Origin and Goals, veröffentl. vom W3C am 29. September 2006. – <<http://www.w3.org/TR/2006/REC-xml-20060816/#sec-origin-goals>>

* Alle aufgeführten Webseiten wurden am 18.12.2013 abgerufen.

Vom Zeichen zur Schrift.

Mit Mustererkennung zur automatisierten Schreiberhanderkennung in mittelalterlichen und frühneuzeitlichen Handschriften

"Deutschland befindet sich in einer Phase intensiv betriebener und mit einem hohen finanziellen Aufwand verbundener Digitalisierung seiner historischen Bestände. Für die Mediävistik und Frühneuzeitforschung stellt hierbei die Digitalisierung der dem Mittelalter und der Renaissance entstammenden Hss. ein zentrales Feld dar."¹ Die Nutzung der Digitalisate allein als digitale Lesekopie durch den betrachtenden Forscher würde das Erkenntnispotential, das dem Digitalisat selbst innewohnt, schlechterdings vergeuden.

Die Gewinnung und Nutzbarmachung der digital vorhandenen Informationen, welche die Analyse des dem Digitalisat zugrunde liegenden materiellen Objekts flankieren und sinnvoll ergänzen können, ist eine der ältesten Fragen digitaler Geisteswissenschaften. Eine zentrale Anwendung ist die Optical Character Recognition (OCR oder automatisierte Texterkennung), die der Herstellung eines maschinenlesbaren Textes aus bildhaft vorliegender Information dient. Sie stößt allerdings bei der Verarbeitung von Handschriftendigitalisaten an ihre Grenzen, was sich anhand des formalen Ablaufs einer Texterkennung veranschaulichen lässt:

1. Anfertigen eines Digitalisates in geeigneter Qualität gegebenenfalls Bildkorrekturen wie das Geraderücken schräg aufgenommener Seiten, Glättung von Rundungen aufgrund von Materialbiegung, etc.
2. Binarisierung der Farbwerte zur deutlichen Trennung von Schrift und Hintergrund
3. Segmentierung der Schrift, z.B. in Linien und Einzelworte
4. Mustererkennung, bei der zu erkennende Formen mit einem vorhandenen Zeichenvorrat verglichen werden
5. Im Falle der Übereinstimmung Zuweisung eines Zeichencodes nach üblicher Textkodierung (=UTF-8)

Die eigentliche Herausforderung an die OCR für Handschriften liegt in Arbeitsschritt 5, da hier zunächst durch langwieriges (und meist manuelles) Erstellen eines Zeichenvorrat, die sogenannte 'ground truth', angelegt werden muss, auf deren

¹ Thomas Haye und Stephan Müller: Mittelalter-Philologie im Internet. 38. Beitrag: Digitalisierung mittelalterlicher Handschriften aus Sicht der Forschung, in: Zeitschrift für deutsches Altertum und deutsche Literatur 140 (2011), S. 416–420, hier S. 416.

Grundlage des Training eines Klassifikators erfolgt. Dieser Klassifikator ist Kern des OCR-Systems und ermöglicht es, Muster (pattern) in der Vorlage Zeichen aus dem Zeichenvorrat zuzuordnen. Diese auf einer festgesetzten Wahrheit beruhende Beziehung zwischen den Bildmustern und den kodierten Zeichen ist nur für die Typen von Mustern gültig, die zum Training verwendet wurden, also z. B. für eine Schrifttype oder eine bestimmte Schriftform. Da die Handschrift jedes Schreibers eigene Charakteristika aufweist, welche sie zumindest von anderen Händen unterscheidbar macht, können die in den meisten Handschriftendigitalisaten aufgefundenen Muster nicht ohne weiteres eindeutig kodierten Zeichen zugewiesen werden, so dass die automatische Erstellung maschinenlesbaren Textes anhand von Digitalisaten in den allermeisten Fällen zu unbrauchbaren, weil fehlerbehafteten Ergebnissen führt. Sollte es jedoch gelingen, anhand spezifischer Merkmale eine Schreiberhand (unter Abstraktion von nicht mehr quantifizierbaren Abweichungsquellen wie Lebensalter und Tagesform des Amanuensis oder dem Zustand der Schreibmaterialien und -utensilien) von anderen Schreiberhänden abzugrenzen, so könnte damit die wichtige Fragestellung nach dem Schreiber automatisiert werden.

In unserem Projekt haben wir also das Untersuchungsziel umgekehrt: Als Ergebnis der Analyse von Handschriftenabbildungen steht nicht ein elektronischer Text, sondern die Identifikation der schreibenden Hand bzw. Hände. Eine mögliche Vorgehensweise zur Lösung dieser Aufgabe ist es, anhand einer von einem sicher zugeordneten Schreiber angefertigten Handschrift die Charakteristika dieser Schrift als 'ground truth' anzutrainieren. Hierzu gehören Buchstabengröße und –abstand, Dichte des Schriftbildes, Neigung u. a., aber nicht notwendig, wie in der klassischen, vom forschenden menschlichen Auge ausgehenden Paläographie, einzelne Buchstabenformen. Basierend auf diesen Charakteristika wird überprüft, ob es möglich ist, diese Hand in anderen Handschriften nachzuweisen. Die Erkennungsgenauigkeit muss dazu aufgrund des Trainings mit einer Handschrift in einem anderen Codex über einem zu definierenden Schwellwert (threshold) liegen, um als Indiz gewertet zu werden, dass derselbe Schreiber die Handschrift geschrieben haben könnte. Der angestrebte Algorithmus würde damit sonstige paläographische oder kodikologische Befunde unterstützende bzw. ergänzende Argumente zur Verifikation von unsicheren Zuschreibungen liefern. Im Gegenzug müsste das Unterschreiten dieses Schwellwertes Argumente für Falsifikationen solcher Zuschreibungen ermöglichen. Darüber hinaus sollte es bei entsprechender Materialbasis und angepassten Schwellwerten möglich sein, Schriftfamilien zu unterscheiden. Würde man die Eigenschaften der wichtigsten Schriftfamilien und räumliche Zusammenhänge, wie etwa insulare Schriftformen in entsprechenden Zeichenvorräten sammeln, wäre es denkbar, Hinweise auf die Datierung und Lokalisierung einer Handschrift aus diesen Vergleichsdaten zu ermitteln.

Ein alternativer Ansatz wäre eine Betrachtung des Schriftbildes als Ganzes und die Bestimmung der Ähnlichkeit von Texten. So könnte ein Grad der Ähnlichkeit eine

Aussage zulassen, ob ein Text eher vom selben oder von einem anderen Schreiber stammt.

Um die aufgestellten Thesen über die Nützlichkeit solcher Erkennungsalgorithmen zu testen, musste zunächst passendes digitalisiertes Material ausgewählt werden. Dabei fiel die Entscheidung auf die wichtigste frühmittelalterliche Buchschrift, die karolingischen Minuskel, die sich durch einheitliche, relativ stark standardisierte Formen und eine meist geringe individuelle Varianz auszeichnet. Diese allgemeinen Spezifika unterstützen die Brauchbarkeit von Digitalisten karolingischer Handschriften ebenso wie deren meist hohes kodikologisches Niveau, das bei entsprechender Fotoaustattung nur wenig Nachbearbeitung erforderlich macht. Mit den im Rahmen des "Europeana Regia"-Projekts digitalisierten Codices Weissenburgenses der HAB steht eine breite Materialbasis zur Verfügung, die weitere Vorteile aufweist: Die weitaus meisten Codices stammen aus dem Skriptorium des Klosters Weißenburg im Elsass, sind also regional und zeitlich gut einzuordnen. Mit den DFG-Richtlinien entsprechenden Katalog von Hans Butzmann² sind diese Handschriften außerdem kodikologisch gut erschlossen. Die Beschreibungen liefern die Vorlagen der Schreiberidentifikation, die es zu verifizieren (oder falsifizieren) galt.

Aus diesem Bestand haben wir in einem ersten Schritt eine Handschrift gewählt, die von einem identifizierten und uns bekannten Schreiber geschrieben worden ist. Die Entscheidung fiel auf Cod. Guelf. 62 Weiss., der nach den Angaben des Kolophons zwischen 819 und 826 von dem Mönch Waldmann während des Abbaatiats von Gerhoh geschrieben worden ist.³ Auf diese Handschrift wurde der Algorithmus angewendet. Weiterhin musste es eine zweite Handschrift geben, die teilweise, aber nicht vollständig von demselben Schreiber angefertigt worden ist, um festzustellen, ob das Verfahren die Handwechsel und damit die Anteile der einzelnen Schreiber erkennt. Hier bot sich Cod. Guelf. 63 Weiss. an, an dem wiederum Waldmann selbst mit zwei weiteren Kopisten, vielleicht seinen Schülern, gearbeitet hat.⁴

Ansätze

In unseren bisherigen Arbeiten haben wir uns zunächst überwiegend mit dem alternativen Ansatz beschäftigt, bei dem die Merkmale zur Schreibererkennung von dem Textblock eines Blattes abgenommen werden. Wie erwähnt, wird nicht versucht, einzelne Zeichen als pattern zu identifizieren, sondern die Handschriftenseite wird als Ganzes interpretiert. Der hochdimensionale Merkmalsvektor einer Seite wird dann mit anderen Seiten verglichen. Je ähnlicher die Merkmalsvektoren sind, desto wahrscheinlicher ist es, dass diese von der selben Hand stammen. Um die Extraktion

² Hans Butzmann: Die Weissenburger Handschriften, Frankfurt am Main 1964 (Kataloge der Herzog August Bibliothek Wolfenbüttel: Neue Reihe, Bd. 10)

³ Ebd., S. 202.

⁴ Ebd., S. 203.

des Textes möglichst einfach zu gestalten, wurden zunächst „typische“ Seiten, vor allem solche ohne Buchschmuck, ausgewählt und für die Gewinnung der Eigenschaften zu Grunde gelegt.

Bei geeigneter Auswahl von Merkmalen konnte gezeigt werden, dass der Schreiber der Handschrift 62 Weiss. auch in 63 Weiss. tätig war. Es konnten die Seiten auf denen er schreibt von denjenigen Seiten geschieden werden, auf denen andere Schreiber tätig sind.

Ergebnisse pro Handschrift

Nach diesem erfolgreichen ersten Test wurden die Algorithmen auf weitere Beispiele angewendet. Diese waren die Handschriften Cod. Guelf. 18 Weiss., 10 Weiss. und 14 Weiss. Dort soll es Überschneidungen der Haupthände geben, die in den Katalogen aber nicht näher lokализiert sind.

Zusammenfassung / Ausblick

Die ersten Tests haben gezeigt, dass eine automatisierte Schreibererkennung basierend auf einem hochdimensionalen Merkmalsvektor einer Textseite möglich ist. Umfangreichere Tests sind nötig, um die universelle Anwendbarkeit zu überprüfen. Als Alternative wird parallel dazu an einem Ansatz gearbeitet, der zeilenorientiert auf völlig anderen Merkmalen arbeitet und eine Klassifikation basierend auf trainierten Klassifikatoren ermöglicht.

AUFBAU EINES KORPUS ZUR BEOBECHTUNG DES SCHREIBGEBRAUCHS IM DEUTSCHEN

PETER M. FISCHER

INSTITUT FÜR DEUTSCHE SPRACHE

PETER.FISCHER@IDS-MANNHEIM.DE

JENS KIRSTEN

THÜRINGER BUCHLÖWE – SCHREIBWETTBEWERB DER

LITERARISCHEN GESELLSCHAFT THÜRINGEN

JENSW.KIRSTEN@GMX.DE

ANDREAS WITT

INSTITUT FÜR DEUTSCHE SPRACHE

WITT@IDS-MANNHEIM.DE

Abstract

Mitte 2013 startete das BMBF-geförderte Forschungsprojekt „Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen“. Das Gemeinschaftsprojekt zwischen dem Institut für Deutsche Sprache, dem Seminar für Computerlinguistik an der Universität des Saarlandes, der Duden Sprachtechnologie am Bibliographischen Institut und der Redaktion Brockhaus-Wahrig bei wissenmedia in der inmediaONE] GmbH hat zum Ziel, korpusbasierte Instrumentarien für eine systematische Beobachtung des Schreibgebrauchs im deutschsprachigen Raum zu entwickeln. Die Beobachtung erstreckt sich dabei nicht nur auf den Schreibusus der sog. *professionellen Schreiber* in gedruckten oder elektronischen Medien sondern auch auf den von Schülern und privaten Internetnutzern. Während die beteiligten Projektpartner bereits über umfassende, aktuelle Textkorpora verfügen, wie beispielsweise das Deutsche Referenzkorpus DEREKO (IDS, 2010), das WAHRIG Textkorpus^{digital} (Krome, 2010) oder das Dudenkorpus

(Münzberg, 2011), welche allesamt Zeitungs- und Zeitschriftentexte sowie auch fiktionale und wissenschaftliche Texte aus dem gesamten deutschsprachigen Raum beinhalten, konzentriert sich der weitere Korpusaufbau auf die Miteinbeziehung von Schüleraufsätzen und Internetbeiträgen, die bislang noch gar nicht abgedeckt sind. Für den Aufbau des Gesamtkorpus ergeben sich damit drei große Bereiche bzw. (virtuelle) Subkorpora, die ihre jeweiligen Herausforderungen mit sich bringen und nun näher betrachtet werden sollen.

Professionelle Schreiber

In diesem Subkorpus sollen die bereits vorhandenen Ressourcen, unter ihnen die o.g. drei größten Korpora der deutschen Gegenwartssprache im deutschen Sprachraum, vernetzt und, wo erforderlich, weiter aufbereitet werden. Wir werden zeigen, wie wir zu diesem Zweck sämtliche Primärdaten in ein einheitliches, für die Vergleichsforschung geeignetes Datenformat überführen und durch Annotation mehrerer linguistischer Ebenen (wie POS, Lemmatisierung, partielle Konstituentenstrukturen, orthografische Differenzierung) sowie durch metalinguistische Informationen auf eine differenzierte linguistische Auswertung vorbereiten müssen.

Internetnutzer

Zwar verfügen die Partner auch über Korpora aus internetbasierten Textquellen (z.B. Wikipedia), jedoch liegt der Fokus für diesen Bereich auf weniger kontrollierten Textsorten wie e-Mails, Weblogs (inklusive Mikroblogs wie Twitter) sowie Texte aus Diskussionsforen. Hierzu ist zum einen die rechtliche Situation zu klären, d.h., ob und welche Art von Lizenzen für welche Textsorte von wem (z.B. nur Forenbetreiber oder auch Autoren?) zu erwerben sind, ob und inwieweit die Texte (möglicherweise nicht nur ihre Metadaten) zu anonymisieren sind. Zum zweiten müssen die Betreiber von Foren, Blogs, e-Mail-Archiven und Mailinglisten (darunter Privatpersonen, Verlage und Firmen) kontaktiert werden, um Lizenzen je nach den rechtlichen Vorgaben sowie nach Möglichkeit Archivversionen fortlaufend erwerben zu können. Zum dritten ergeben sich für die Basisannotation (Struktur und Metadaten) der Texte besondere Anforderungen, die gängige Korpus-Textmodelle, die

vorwiegend auf Zeitungstexte, Belletristik und Fachtexte abzielen, noch nicht abdecken, und daher umgesetzt werden müssen.

Schüler

Für diesen Bereich werden Schülertexte unterschiedlicher Art akquiriert und in die Korpora integriert. Dafür werden zunächst Kooperationen mit Schulen eingegangen, um Kopien von Schülertexten, Klausurarbeiten und Abituraufsätzen zu bekommen. Automatische Verfahren unterstützen die Digitalisierung dieser Texte. Des Weiteren werden Textdaten in digitaler Form von Schülern in Kooperation mit Schulen direkt erhoben. Schließlich werden Kooperationen mit Literaturwettbewerben für Schüler eingegangen, um Texte, die in digitaler Form eingereicht wurden, zu erwerben. Dies werden wir an der Zusammenarbeit mit der Literarischen Gesellschaft Thüringen e.V. exemplifizieren, die jährlich den Wettbewerb „Thüringer Buchlöwe – Schreibwettbewerb der Literarischen Gesellschaft Thüringen“ ausrichtet.

Gesamtkorpus

Die akquirierten und digitalisierten Texte werden konvertiert und nach einem TEI-basierten Textmodell aufbereitet und annotiert. Im Falle der Schülertexte sind rechtliche Fragen bzgl. Möglichkeiten der Herausgabe von z.B. Abituraufsätzen durch Schulen und bzgl. Anonymisierungspflicht im Vorfeld zu klären. Der Gesamtaufbau besteht im Ergebnis also in der Zusammentragung eines außerordentlich großen virtuellen Korpus zur deutschen Gegenwartssprache, das aus zu unterschiedlichen Bedingungen lizenziertem Material besteht. Das virtuelle Korpus wird für ausgesuchte Zwecke, insbesondere die Beobachtung der Schreibverwendung, genutzt werden können; eine Erweiterung der Nutzungsszenarien unter Wahrung der Lizenzanforderungen ist auch möglich.

Referenzen

Belica, C., Kupietz, M., Lüngen, H., Witt, A. (2010): The morphosyntactic annotation of DEREKO: Interpretation, opportunities and pitfalls. In M. Konopka, J. Kubczak, C. Mair, F. Šticha, and U. Wassner, editors, Selected

contributions from the conference Grammar and Corpora 2009, im Druck in Tübingen. Gunter Narr Verlag.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., Pinkal, M. (2006): The SALSA Corpus: a German corpus resource for lexical semantics. LREC 2006

IDS (2010): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2010-I. Institut für Deutsche Sprache. Mannheim. <http://www.ids-mannheim.de/kl/projekte/archiv.html>

Krome, S. (2010): Die deutsche Gegenwartssprache im Fokus korpusbasierter Lexikographie. Korpora als Grundlage moderner allgemeinsprachlicher Wörterbücher am Beispiel des WAHRIG Textkorpus^{digital}. In: I. Kratochvílová, N. R. Wolf (Hgg.): Kompendium Korpuslinguistik. Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive. Universitätsverlag Winter, Heidelberg 2010, S. 117-134.

Kupietz, M., Witt, A., Belica, C., Keibel, H. (2010): The German Reference Corpus DEREKO: A Primordial Sample for Linguistic Research. In: LREC 2010 Main Conference Proceedings. Malta.

Lemnitzer, L., Geyken, A., Beißwenger, M., Storrer, A. (2011): CMC as a component of a balanced, TEI-encoded corpus representing contemporary German: goals, motivation, design issues. Abstract eines Papers zur Präsentation auf dem TEI Members' Meeting, Würzburg, Oktober 2011.

Münzberg, Franziska (2011): Korpusrecherche in der Dudenredaktion. Ein Werkstattbericht. In: Marek Konopka et al. (Hg.): Grammatik und Korpora 2009. Tübingen: Narr Francke Attempt 2011, 181–197.

The TEI Consortium (2007): Guidelines for Electronic Text Encoding and Interchange (TEI P5). The TEI Consortium. <http://www.tei-c.org/Guidelines/P5/>

Walter, S., Pinkal, M. (2005): Computational Linguistic Support for Legal Ontology Construction. In Proceedings of ICAIL 2005

Witt, A., Kupietz, M., Keibel, H. (2009): DEREKO goes P5: Customizing TEI P5 for the Mannheim German Reference Corpus, Micropaper IN: Text encoding in the era of mass digitization, Online-Konferenz-Proceedings des TEI Members Meetings 2009, <http://www.lib.umich.edu/spo/teimeeting09/>

Max Weber Stiftung Rheinallee 6 53173 Bonn Germany

Wissenschaftliches Bloggen bei de.hypotheses.org

Wissenschaftliche Blogs sind ein Instrument für die wissenschaftliche Arbeit. Zur Wissenschaftskommunikation mit Fachgemeinschaft und Öffentlichkeit, für Veranstaltungankündigungen und Tagungsberichte, als Zettelkasten, Vernetzungsoffer und Diskussionsort können Blogs dienen. Wissenschaftliche Texte können schnell und einfach digital publiziert und kommentiert werden.

Der Workshop bietet eine Einführung in konzeptionelle und praktische Aspekte des Bloggens. In einer kurzen Einführung geht es um Fragen wie: Was macht einen guten Blog aus? Welche Sprache ist angemessen? Welche Regeln existieren für das wissenschaftliche Bloggen?

Im Hauptteil kann jeder Teilnehmer in seinem eigenen Schulungsblog von de.hypotheses.org mit Wordpress Bloggen intensiv praktisch üben. Immer wieder gibt es dabei Exkurse zu rechtlichen Grundlagen (Urheberrechte, OpenAccess, CreativeCommons Impressum) und Tipps für die Startphase eines wissenschaftlichen Blogs.

Teilnehmer

bis max. 30 Teilnehmer/innen

Technische Ausstattung

Beamer & Leinwand

Internetverbindung

Eigener Laptop

Ausreichend Steckdosen

Deutsches Forum für
Kunstgeschichte Paris

Deutsches Historisches
Institut London

Deutsches Historisches
Institut Moskau

Deutsches Historisches
Institut Paris

Deutsches Historisches
Institut Rom

Deutsches Historisches
Institut Warschau

Deutsches Historisches
Institut Washington D.C.

Deutsches Institut
für Japanstudien Tokyo

Orient-Institut Beirut

Orient-Institut Istanbul

Bundesunmittelbare Stiftung
des öffentlichen Rechts

In this poster we present the CLLD (Cross-Linguistic Linked Data) project [1] - an MPG-internally funded project aimed at helping to preserve the world's language diversity heritage. The particular problem this project sets out to solve is bridging the gap between data collection and data sharing.

This problem can be framed with two observations from this year's ALT (Association of Linguistic Typology) 10 Conference [2]:

- Most talks about results reached with quantitative methods have been based on the WALS [3] dataset.
- Many linguists at ALT 10 work on and with "private" databases, often enriching or expanding the WALS dataset; but basically none of these are accessible or published.

A similar gap exists for dictionaries or wordlists for little known languages:

Collecting these is highly valued within the community, but publication via traditional channels is basically impossible because there are too few readers for such dictionaries to make publication as books worthwhile.

The CLLD project aims at solving this problem by providing sustainable, interoperable data publication structures, thereby offering incentives and lowering the barriers to publish linguistic databases. This strategy is implemented as follows:

1. The project develops a web application framework tailored for the publication of lexical or typological linguistic data - a CMS for linguistic databases [4].
2. Based on this framework, the databases of the department of linguistics at MPI EVA [5] are published. In addition, a language catalog (Glottolog [6]) is published, which allows cross-referencing language resources across distributed databases.
3. Two online database journals (one for dictionaries and one for typological databases) will be started, allowing for submission of databases which are too small to be published as stand-alone resources.
4. Data is published following Linked Data principles in order to allow for
 - uniform data access and interoperability of distributed resources at web-scale,
 - a well-defined lower boundary of service in a "graceful degradation of service" scenario (see below).
5. A strategy for "graceful degradation of service" is adopted, helping to describe service levels and requirements thereby making transitions of ownership/maintenance of services easier.

After the first year of the 4-year funding period, the following results have been obtained:

- The clld software package is used to publish the typological databases WALS, APiCS, eWAVE, PHOIBLE, AFBO as well as the lexical database WOLD and the language catalog Glottolog (see [1] and [7]); thus validating the concept and data models underlying its design.
- The process to gather initial submissions for the two database journals has been started.
- Applicability of the software and the process to linguistic resources outside the MPI EVA has been shown with the publication of the Phonetics Information Base and Lexicon (PHOIBLE).

[1] <http://clld.org>

[2] http://www.eva.mpg.de/lingua/conference/2013_ALT10/

[3] The World Atlas Of Language Structures Online. <http://wals.info>

[4] <https://github.com/clld/clld>

[5] <http://www.eva.mpg.de/linguistics/index.html>

[6] Glottolog. <http://glottolog.org>

Themenbereich: Vom analytischen Mehrwert digitaler Werkzeuge für die Geisteswissenschaften

Titel: Was bringt die Anwendung von QDA-Software für die Analyse qualitativer Daten

Autor: Susanne Friese, Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften, Göttingen

Kontakt: friese@mmg.mpg.de

Abstract: Erste Experimente Computer in der qualitativen Datenanalyse (QDA) einzusetzen, fanden Anfang der 80er Jahre statt. Damals versuchte man sich Datenmanagement-, inhaltsanalytische und Textverarbeitungsprogramme für diesen Zweck zu Nutzen zu machen. Bis zu diesem Zeitpunkt waren Farbstifte, Karteikästen, Schere, Kleber und manchmal auch die Stricknadeln der damals üblichen Randlochkartensysteme wesentliche Handwerkszeuge des qualitativen Forschers. Mit der Schere wurden wichtige Textstellen ausgeschnitten und mit inhaltlich ähnlichen Textpassagen zusammen auf ein Blatt Papier geklebt oder in Karteikästen und Schuhkartons eingesortiert. Farbstifte dienten zur Markierung und zur Feingliederung der als wichtig erachteten Textpassagen.

Erste Computerprogramme zur direkten Unterstützung der qualitativen Datenanalyse ahmten die manuelle Vorgehensweise nach und ersetzten somit Schere, Farbstifte Klebstoff und Karteikarten und Wände mit post-it Notizen. Das Hauptaugenmerk war auf das Kodieren (= Ausschneiden und Ablage unter einem geeigneten Begriff) von Textpassagen und das spätere Wiederfinden (= Suche in Karteikästen) gerichtet. Diese erste Generation von QDA Programmen hieß folgerichtig "Code & Retrieve" Software. Dazu gehörten die ersten Versionen von THE ETHNOGRAPH, Qualpro und TAP (Fielding & Lee, 1998). Seit Ende der 1980er bis heute wurden eine Reihe von weiteren Programmen u.a. in Deutschland (AQUAD, ATLAS.ti, MAXQDA), Holland (Qualitan), Dänemark (Textbase Alpha), England (Digital Replay System), Nordamerika (Transana, QDA-Minder, Dedoose) und Australien (Nud*ist/NVivo) entwickelt.

Trotz der weltweiten Verbreitung und Nutzung von QDA Software, wird dem Computer als Hilfsmittel für die qualitative Datenanalyse zum Teil heute immer noch mit Misstrauen begegnet. Ein Grund dafür ist, dass die Anwendung von Software zur Datenauswertung eher

dem quantitativen Forschungsparadigma zugeordnet wird und daher der Einsatz in der QDA epistemologisch bedenklich erscheint (Tesch, 1990; Friese, 2011). Doktoranden müssen zum Teil noch immer gegenüber ihren Betreuer argumentieren, warum Sie ihre Daten softwaregestützt, mit den dafür vorgesehenen CAQDAS Tools¹, und nicht manuell oder mit Word oder Excel auswerten. Damit im Zusammenhang steht auch die Anforderung, die an die Analyse von qualitativen „weichen“ Daten gestellt wird. Sie ist deutlich geringer, als die für quantitative „harte“ Daten. Für die Auswertung Letzterer werden immer anspruchsvollere Verfahren verlangt. Eine Anfrage, ob nicht auch eine Bearbeitung in Excel ausreichend wäre, würde wahrscheinlich sofort abgelehnt. Einfache Kreuztabellen und Chi-Quadrat-Tests reichen für eine Doktorarbeit nicht. Wenn es hingegen um die Auswertung qualitativer Daten geht, genügt manchmal schon der Hinweis im Methodenteil, dass man die Daten genauesten gelesen hat (Stichwort: *close reading of the data*).

Es gibt zwei Gründe für diese Situation. Der erste Grund ist, dass die Programme zur Auswertung qualitativer Daten nicht beherrscht werden und man auch nicht die Zeit investieren möchte, diese zu lernen. Dies ist zum Teil ein Generationenproblem, dass sich von alleine lösen wird. Zum anderen ist es aber auch bedingt durch die fehlenden Ausbildungsmöglichkeiten. Das Erlernen von CAQDAS steht auf den wenigsten Lehrplänen und man muss es sich selbst aneignen oder die Teilnahme an einem Workshops selbst finanzieren. Darüber hinaus gibt es oftmals keine weitere Betreuung. Der zweite Grund ist eine generelle Ablehnungshaltung. In manchen Fällen ist eine solche Ablehnung angebracht und zwar immer dann, wenn der methodische Ansatz nicht davon ausgeht, dass die Daten kodiert werden (z.B. bei der Anwendung Sequenzanalytische Verfahren oder in der Objektiven Hermeneutik). In anderen Fällen ist die negative Haltung zu Software bedingt durch die Angst, dass der Forscher die Nähe zu den Daten verlieren würde, oder dass die Software die Daten automatisch kodieren und somit interpretieren würde, was natürlich keine Maschine leisten kann. Ein weiteres Argument gegen den Einsatz von Software ist, dass Software zu oberflächlichen Analysen führen würde. Dies alles sind Vorurteile, die auf einem unvollständigen Verständnis beruhen, was Software leisten bzw. nicht leisten kann. Leider gibt es natürlich auch jede Menge schlecht durchgeführte „quick & dirty“ Analysen, was aber nicht der Software an sich angelastet werden kann. Eine weitere Barriere zum Abbau der

¹ Das Akronym CAQDAS wurde von den Leitern des gleichnamigen Projektes an der University of Surrey, Guildford, UK, entwickelt. Es steht für „Computer Assisted Qualitative Data Analysis Software“.

Vorurteilen ist, dass in Forschungsberichten zumeist nur geschrieben wird, dass Software angewendet wurde und nicht wie und wozu (Paulus et al., 2014).

Der Sinn und Zweck von CAQDAS, wie der Name schon sagt, ist nicht Daten zu analysieren. Sondern es ist nur ein Tool, das den Analyseprozess unterstützt. Diese Erklärung wird von Einigen allerdings wiederum als Argument gegen die Anwendung von Software angeführt, gemäß der Logik: „Wenn die Software die Daten nicht kodiert, wofür ist sie dann zu gebrauchen?“ Und ohne eines der CAQDAS Pakete jemals richtig ausprobiert zu haben, fällen sie das Urteil, dass Software im qualitative Forschungskontext nichts zu suchen hat und kehren zurück zu ihren manuellen Methoden, Scheren, Klebern, bunten Stiften, Karteikästen und großen Wandflächen. Hier ein online Zitat als Anleitung für die Analyse ethnographischer Daten aus dem Jahre 2012:

“Complex software can get in the way of analysis. Sometimes, all you need to do is print out your fieldnotes, take a highlighter, and talk it over with yourself or with someone else.

Big walls are important – I've found that I need lots of wall space to spread my ideas out. Carry sticky notes with you everywhere you go.....”²

Basierend auf meiner langjährigen Erfahrung mit CAQDAS, ist meine Antwort darauf, dass sich ForscherInnen mit dieser Einstellung Möglichkeiten nehmen, auch mit Hinblick auf die Validität ihrer Forschung. Mit Hilfe von Software wird es einfacher, bzw. erst möglich auch größere Datenmengen zu analysieren. Ein zusätzliches Plus ist, dass eine sauber durchgeführte softwaregestützte Analyse die Güte der Forschungsergebnisse erhöht. Insbesondere, wenn man die konzeptuelle Phase des Analyseprozesses erreicht hat, vergisst man leicht die Inhalte, die hinter den Konzepten stehen. Testet man verschiedene Modelle um zu sehen, wie alles zusammen passen könnte, ist es sehr wichtig immer wieder einen Blick auf die Rohdaten zu werfen. Ist dies mit viel Handarbeit verbunden, unterbleibt dieser Schritt oftmals; bzw. man stützt sich nur auf ein paar Textstellen, weil auf den Notizzetteln an der Wand nur ein paar der „guten“ Zitate vermerkt sind und das Durchsehen aller Feldnotizen zu aufwendig wäre. Ist das Ausgangsmaterial nur ein paar Mausklicks entfernt, wird man eher

² <http://ethnographymatters.net/2012/09/04/the-tools-we-use-gahhhh-where-is-the-killer-qualitative-analysis-app/>

mal nachschauen, ob nicht nur das Ergebnis gut aussieht und sich erklären lässt, sondern ob auch alle dazu beitragenden Datensegmente zur Erklärung und Interpretation passen.

Wenn Software *richtig* angewendet wird, dann bietet sie die Möglichkeiten Ideen, Hypothesen, theoretische Konstrukte oder Modelle in jeder Phase des Analyseprozesses zu verifizieren oder zu falsifizieren und zwar genauer, als dies in einer händischen Analyse je möglich ist. Das Wort „richtig“ ist hierbei jedoch das ausschlaggebende Wort. In vielen Büchern, die Auswertungsverfahren für qualitative Daten beschreiben, findet man oftmals nur einen Zusatz, dass man zur Umsetzung der Vorgehensweise heutzutage auch Software verwenden kann. Es wird aber nicht beschrieben, wie. Das scheint selbsterklärend zu sein. In Büchern und Aufsätzen, die die verschiedenen Tools beschreiben, werden wiederum zumeist nur die Softwarefunktionen und die Mausklicks erläutert. Der Brückenschlag zur methodischen Vorgehensweise wird selten gemacht. Eine Ausnahme ist das Buch von Gibbs (2007). Er erklärt die Anwendung bestimmter Funktionen der Software NVivo für unterschiedliche analytische Stile wie z. B. Grounded Theory, Narrative Analyse, Biographieforschung oder Diskursanalyse. Mein Ansatz, die NCT Methode der computergestützten qualitativen Datenanalyse, geht noch einen Schritt weiter, bzw. einen anderen Weg (Friese, 2012/2014). Die Methode basiert auf über 20 Jahren Erfahrung in der Durchführung von softwaregestützten Analysen, Projektberatungen und dem Unterrichten von Software. Insbesondere die Fehlversuche, inklusive meiner eigenen, und Projekte bei denen die Nutzer ausgestiegen sind um mit dem Altbewährten, sei es auf Papier oder in Excel weiter zu machen, haben dazu geführt einen Weg zu beschreiben, wie man effizient ein Projekt aufbaut und daraus resultierend Analysen durchführen kann, die einen Mehrwert bieten. Auf der untersten Ebene hilft CAQDAS wahrscheinlich vielen Anwendern ihre Daten zu verwalten und auf simple Weise mit Hilfe von Kodes zu ordnen. Darüber hinaus bietet sie aber noch viel mehr Funktionalität und diesen erweiterten Mehrwert in Bezug auf analytische Möglichkeiten möchte ich im Vortrag (bzw. schriftlichen Beitrag) anhand von Beispielen präsentieren. Einführend gehe ich kurz auf den allgemeinen Mehrwert von digitaler Analyse auf der untersten Ebene mit Bezug auf die Datenverwaltung und Kodierung ein. Dies ist jedoch nicht neu und findet sich in der einschlägigen Literatur. Vielmehr möchte ich anhand von kodierten CAQDAS Projekten aufzeigen, welche weiterführenden Analysen mit Hilfe digitaler Tools möglich sind, die ohne diese Tools nicht durchgeführt werden könnten. Des Weiteren stelle ich die Auswirkungen eines schlechten oder nicht adäquaten Projektdesigns anhand verschiedener Versionen eines Projekts dar, um zu zeigen welche Auswirkung das

Design und der Projektaufbau auf den Zusatznutzen haben. Diese Beispiele können Anfängern aufzeigen, was zu beachten ist, damit man die zur Verfügung stehenden Tools von Anfang an effizient nutzt; und aktuellen Anwendern, warum sie an bestimmten Punkten mit ihrer Analyse nicht weiter kommen oder das Gefühl haben, dass manche der Auswertungstools für sie nicht nutzbar sind.

Literatur:

- Fielding, Nigel G. and Lee, Raymond M. (1998). Computer Analysis and Qualitative Research. London: Sage.
- Friese, Susanne (2012). Qualitative Data Analysis with ATLAS.ti. London: Sage. (2. Ausgabe März 2014)
- Friese, Susanne (2011). Computergestützte Analyse qualitativer Daten. In: R. Ayaß und J. Bergmann (Hrsg.), Sammelband: Qualitative Methoden der Medienforschung. Mannheim: Verlag für Gesprächsforschung. Online: <http://www.verlag-gespraechsforschung.de/2011/ayass.htm>.
- Gibbs, Graham (2007). Qualitative Data Analysis: Exploration with NVivo.
- Paulus, Trena, Woods, Megan, Atkins, David and Rob Macklin, Rob (2014, forthcoming). Current Reporting Practices of ATLAS.ti User in Published Research Studies, in: Friese, Susanne und Ringmayr, Thomas (Hrsg.), ATLAS.ti User Conference 2013: Fostering Dialog on Qualitative Methods. University Press, Technical University Berlin. <http://nbn-resolving.de/urn:nbn:de:kobv:83-opus4-44295>
- Tesch, Renata (1990). Qualitative Research: Analysis Types and Software Tools. New York (Falmer).

Die Öffnung bibliothekarischer Daten - Eine Spielwiese

Seit der ersten Freigabe von Katalogdaten durch das hbz im Jahr 2010 haben mittlerweile fast alle großen Bibliotheksverbünde sowie die Deutsche Nationalbibliothek nachgezogen.

Die meisten dieser Daten werden als (Linked) Open Data unter der Creative Commons Zero Lizenz zur Verfügung gestellt, d.h. die Daten sind gemeinfrei, sie gehören allen und dürfen zu beliebigen Zwecken und ohne Auflagen genutzt werden.

Es scheint mir aber noch wenige Services, Portale oder Tools zu geben, die sich mit diesen Daten und ihrer Nutzung beschäftigen, und ich habe den Eindruck, dass die Phase des Kennenlernens und Spielens mit den Möglichkeiten erst begonnen hat.

Von den veröffentlichten Institutionen wird gerne auf die Möglichkeiten hingewiesen, die im RDF Format liegen, in welchem die meisten Daten zum Download angeboten werden, nämlich die Adressierung und Einbindung im Rahmen des World Wide Web. Das wird auch bereits gemacht, z.B. von der Wikipedia.

Der Vortrag will, an einem Beispiel, nur ein paar Gedanken dazu beitragen.

Schon im Studium hat mich fasziniert, dass neben der Geschichte der großen deutschen Literaturwerke und ihrer Autoren diejenige der Übersetzungen und Übersetzer ein recht kärgliches Schattendasein gefristet hat. Abgesehen von den Fällen, wo ein literaturgeschichtlich relevanter deutscher Autor zugleich als Übersetzer tätig war, sind mir keine kritischen Werkeditionen bekannt, die Übersetzungen zum Gegenstand haben. So wird es wohl irgendwann bestimmt eine kritische Edition der Tieckschen Don Quixote-Übertragung geben, jedoch so rasch keine der Braufelsschen, obwohl beide eine ähnliche Wirkung entfaltet haben. Die bibliografische Lage war zu meiner Studienzeit ebenfalls mager. Mittlerweile gibt es aber schon eine ganze Reihe Bibliografien zu verschiedenen Nationalliteraturen in deutscher Übersetzung. So z.B. die in einem DFG-Projekt entstandene *Bibliographie niederländischer Literatur (von den Anfängen bis 1830) in deutscher Übersetzung* unter der Leitung von Prof. Dr. Konst an der Freien Universität Berlin, um nur eine zu nennen.

Ohne auch in der Kunst des Bibliografierens ein Experte zu sein, kann ich mir vorstellen, dass es ein mühevolleres Unterfangen ist, überhaupt erst einmal eine Basis für ein solches Unterfangen zu schaffen.

Die jetzt freigegebenen Titeldaten der deutschen Bibliotheksverbünde könnten für das Projekt der *Weltliteratur in deutschen Übersetzungen* jedoch äußerst hilfreich sein.

Folgende Bestände können bereits bezogen und verarbeitet werden:

DNB - Deutsche Nationalbibliothek (ca. 11 Mio. Titeldatensätze)

BVB - Bibliotheksverbund Bayern und KOBV - Kooperativer Bibliotheksverbund Berlin-Brandenburg (ca. 23 Mio. Titeldatensätze)

GBV - Bibliotheksverbund der Länder Bremen, Hamburg, Mecklenburg-Vorpommern, Niedersachsen, Sachsen-Anhalt, Schleswig-Holstein, Thüringen und der Stiftung Preußischer Kulturbesitz (ca. 37 Mio. Titeldatensätze; nur ausgewählte Datenbestände)

HBZ - Verbundkatalog NRW (ca. 18 Mio. Titeldatensätze)

HeBIS - Hessisches Bibliotheksinformationssystem (ca. 17,5 Mio. Titeldatensätze)

SWB - Verbundkatalog des Südwestdeutschen Bibliotheksverbund (ca. 17,5 Mio. Titeldatensätze)

Die Suche nach den relevanten Übersetzungen war bislang nicht so einfach. Über die OPAC-Services der Bibliotheken und Verbünde konnte - auch in den erweiterten Suchen - nur auf ausgewählte Datenfelder der MAB- oder MARC-Datensätze zugegriffen werden. Zugriffe per API waren erlaubnispflichtig und auf spezielle Nutzergruppen beschränkt.

Autorennamen konnten dann freilich in vielen Schreibweisen vorkommen und die Information, dass es sich um eine Übertragung ins Deutsche handelte, konnte auf unterschiedlichste Weise in den vielen Datenfeldern verborgen sein.

Erst seit den 90er Jahren wurde beispielsweise der dreistellige Sprachencode im Feld 37b eines MAB-Titeldatensatzes erfasst - und auch dann keineswegs immer.

Die häufig verwendete Zeichenfolge "[Übers.]" konnte in den MAB-Feldern 100-104 einen Hinweis liefern.

Zeichenfolgen wie "Übers. aus dem", "Übers.:", "Deutsch von", "In der Übersetzung von", "durch ... verdolmetscht" usw. konnten in verschiedensten Feldern eines Titeldatensatzes auftauchen.

Ebenso das Kürzel "<dt.>" im Einheitssachtitel oder einer Nebeneintragung.

Die freien Daten ermöglichen nun aber das Schreiben eigener Filter.

Zudem ist auch die Gemeinsame Normdatei, welche die verschiedenen Ansetzungsformen von Personennamen liefert, mittlerweile frei verfügbar.

Will man nun die in allen deutschen Bibliotheken zu einem Autor vorhandenen Titel in deutscher Übersetzung finden, so scannt man die Daten unter allen Ansetzungsformen des Namens sowie den o.g. Kriterien. Einfach - aber bislang so nicht möglich.

Gleiches gilt natürlich für die Suche nach einem Titel, wenn man z.B. die Don Quixote Übersetzungen des 17. und 18. Jahrhunderts zusammenstellen möchte (was freilich heute kein Desiderat mehr ist).

Der kleine Vortrag möchte an dieser Stelle (in seinem Kern) Beispiele der freien Titeldaten, den Aufbau eines Suchscriptes, die dadurch erzielbaren Ergebnisse und mögliche Weiterverarbeitungen zeigen.

Hat man einmal die Möglichkeit geschaffen, die relevanten Datensätze herauszufiltern, sind natürlich weitere Anwendungsmöglichkeiten denkbar. Die RDF-Triple der offenen Daten lassen insbesondere zu, dass jedes Objekt, z.B. eine Titelansetzung oder ein Autoname, selbst zum Subjekt anderer Triple wird, die auf weitere Ressourcen im Web verweisen.

So könnte ein gedachtes Portal nicht nur die bibliografischen Informationen, sondern z.B. auch gleich den Zugriff auf die Digitalisate aus den wachsenden Sammlungen an Retrodigitalisierungen der Bibliotheken bieten. Informationen zu den Autoren, Lebensdaten, Historisches etc.

Unter Auswertung der Druckorte ist ferner eine Landschaft der literarischen Übersetzungstätigkeit in Kartenform denkbar.

Semi-automatische Differenzanalyse von komplexen Textvarianten

André Gießler andre.giessler@informatik.uni-halle.de

Marcus Pöckelmann marcus.poeckelmann@informatik.uni-halle.de

Jörg Ritter joerg.ritter@informatik.uni-halle.de

Einer der Schwerpunkte von Projekten der Editionsphilologie ist die Untersuchung alter Texte mit Mehrfachüberlieferungen sowie die Textgenese. Dabei stellt sich für die beteiligten Wissenschaftler die Aufgabe, die Verbindungen zwischen den einzelnen Textvarianten herauszuarbeiten und dabei Gemeinsamkeiten und Unterschiede zu erkennen. Oft sind große Textmengen der verschiedenen Varianten zu sichten, einander zuzuordnen und detailliert zu vergleichen, um anschließend als Edition präsentiert werden zu können. Bisher erfolgen die bei der Edition anfallenden, teils sehr gleichartigen und zeitaufwändigen Zwischenstufen in Handarbeit, belegen damit wertvolle Arbeitszeit und setzen einen Gesamtüberblick über das Textmaterial voraus. Die Informationstechnologie bietet heute Möglichkeiten, mit denen die Durchführung vieler dieser Schritte zumindest teilautomatisiert werden kann. Den Geisteswissenschaftlern können Werkzeuge zur Verfügung gestellt werden, die ihnen die Arbeit nicht nur wesentlich erleichtern, sondern auch die Fehleranfälligkeit reduzieren und neue Formen der Auswertung eröffnen.

Das hier vorgestellte, vom BMBF geförderte Gemeinschaftsprojekt von Geisteswissenschaftlern und Informatikern mit dem Ziel, Werkzeuge und Methoden zum Textvergleich und zur Erstellung kritischer und genetischer Editionen zu entwickeln. Diese Methoden sollen generisch und damit auf viele Textformen anwendbar sein. Dazu werden zwei Repräsentanten verschiedenartiger Textformen mit ihren Überlieferungen zur Grundlage genommen, für deren Anforderungen und Eigenheiten jeweils zugeschnittene Verfahren entwickelt und evaluiert werden. Dabei werden die Prozesse von der Erkennung und Lemmatisierung der Wörter, das Auffinden sich entsprechender Textstellen, die Herausarbeitung der Unterschiede und Gemeinsamkeiten, bis hin zur Darstellung in einer genetischen Edition abgedeckt. Im späteren Verlauf des Projektes findet die Verallgemeinerung der gewonnenen Erkenntnisse für möglichst viele Textformen statt.

Ein Teil des Projektes betrachtet einen handschriftlichen Lehrbuchtext aus der Zeit des Spätmittelalters, der in Frühneuhochdeutsch verfasst wurde. Unter der Leitung von Hans-Joachim Solms wird die „Wundarznei“ des Heinrich von Pfalzpaint aus dem 15. Jahrhundert in ihren zehn verfügbaren Überlieferungen untersucht. Ziel der Altgermanisten ist hier, die Varianten zu vergleichen und in einer kritischen Edition und einer Online-Edition mit Synopse und Variantenapparat darzustellen. Ausgangspunkt sind die Handschriften, die in einem ersten Arbeitsschritt von den Geisteswissenschaftlern diplomatisch transkribiert wurden. Da diese Texte in der Sprachstufe Frühneuhochdeutsch verfasst wur-

den, gibt es keine einheitliche Graphie, wodurch dieselben Wörter in verschiedenen Überlieferungen deutlich unterschiedlich geschrieben werden und mit jeder weiteren Überlieferung neue Schreibweisen entdeckt werden. Ein Beispiel ist das Wort Pfeil, welches in den Schreibweisen „pfeil“, „pfœil“, „pfejl“, „pffejl“, „pfeyl“, „pfeyl“ auftritt. Bevor ein Textvergleich stattfinden kann, ist somit eine philologische Aufbereitung nötig, bei der die Wörter erkannt und normalisiert werden. Die Normalisierung wird mit der Lemmatisierung jedes einzelnen Wortes erreicht. Für eine möglichst präzise Abbildung einer Handschrift auf eine andere werden die Wörter zusätzlich noch mit Part-Of-Speech-Tags und morphologischen Attributen wie Kasus, Numerus und Genus versehen. Für die Aufgabe der Lemmatisierung und Annotation existieren bereits verschiedene automatisierte Ansätze, die allerdings nicht auf Handschriften aus dem Frühneuhochdeutschen anwendbar sind, da sie nur eine sehr geringe Toleranz für abweichende Schreibweise (oder Schreibfehler) von Wörtern aufweisen. Die unstetige Graphie in den Handschriften der „Wundarznei“ führt bei ihnen zu geringen Trefferquoten in Bezug auf die Korrektheit der von ihnen vorgeschlagenen Annotationen.

Im Projekt wurde das Werkzeug *Lemmano* entwickelt, das einen semiautomatischen Ansatz verfolgt. Es erlaubt die manuelle Annotierung jedes einzelnen Wortes, in dem es zu dem Wort ähnliche Wortformen ableitet, diese mit zugehörigen Annotationen in Lexika sucht und dem Benutzer so Vorschläge für das aktuelle Wort unterbreitet. Ähnlich heißt hier, dass sich die neue Wortform mittels von Altgermanisten erarbeiteten Ersetzungsregeln, die auf Äquivalenzen bestimmter Buchstabenfolgen basieren, aus dem gegebenen Wort ableiten lässt. Im Gegensatz zu automatischen Ansätzen liegt die Entscheidung für das passende Lemma und die passenden morphologischen Daten beim Anwender.

Die Benutzeroberfläche ist intuitiv verständlich gestaltet und auf die Massenverarbeitung ausgerichtet. *Lemmano* hat sich bei der Annotation der Handschriften der „Wundarznei“ durch Germanisten als sehr große Arbeitserleichterung erwiesen. Da es ein webbasiertes Werkzeug ist, können mehrere Nutzer gleichzeitig annotieren und profitieren von den gelernten Eingaben der anderen Nutzer. Bild 1 zeigt den Dialog für eine Wortform.

Nach der Lemmatisierung der Handschriften lassen diese sich nun mit weiteren digitalen Werkzeugen, die im Rahmen des Projektes realisiert werden, detailliert vergleichen. Der nächste Schritt auf dem Weg zu einer Edition mit synoptischer Darstellung ist die Alignierung der Varianten zueinander. Dabei werden sich entsprechende Textstellen identifiziert, gegenübergestellt und mit Auflistung der Unterschiede in einem Variantenapparat präsentiert.

Das zweite Teilprojekt widmet sich einem neuphilologischen Text in Fremdsprache. Unter der Leitung von Thomas Bremer wird die im späten 18. Jahrhundert in Französisch verfasste „*Histoire philosophique et politique des établissements et du commerce des Européens dans les deux indes*“ von Abbé Guillaume Thomas François Raynal untersucht. Sie gilt auf Grund ihrer Thematik, die Auseinandersetzung mit der europäischen Kolonialpolitik dieser Zeit, als bedeutendes Werk der Aufklärung. Nach dem Verbot der Erstauflage von 1770 erschienen zwei weitere Auflagen sowie eine postume Textfassung. Im Rahmen des Projektes soll eine genetische Edition der Lateinamerika-Bände entstehen, die insbesondere durch ihre Interaktivität die Evolution dieses Werkes nachvollziehbar macht. Dabei liegt das Hauptaugenmerk auf einer abschnittsweisen

Lemmano Korpora Texte Dokumente Lexika Lemmalinks Suche Vergleich Nutzer root: Abmeiden

B8-112r18	ten	,	die	do	bolze	sint	alzo	mä	in	den	buchern	wol	findet		
B8-112r19	czu	den	czeichen	.	so	genilet	keiner	vnd	wurde	her	schle				
B8-112r20	cht	durch	die	hut	gelchollen	,	sundern	her	weylz	danne	die				
B8-112r21	kunt	,	die	darczu	gehoret	.	Hir	findelt	u	nü	recht				
B8-112r22	wi	du	die	pfile	czihen	salt	etc	.	Item	,	nu	wil	ich		
B8-112r23	dich	lernen	,	wi	du	die	pfile	czihen	solt	,	sie	haben	holcz		
B8-112r24	in	dem	tulle	ader	nicht	.	Item	,	kompt	dir	einer	czu	vnd		
B8-112r25	stegkit	ein	pfiylen	in	yme	,	so	wil	dich	lernen	,	das	du		
B8-112r26	das	heruf	czuhilt	ane	czangen	.	vnd	bedarf	ouch	keinē					
B8-112r27	sniten	,	war												
B8-112r28	so	ilt	ditt												
B8-112r29	czangen	.													
B8-112r30	so	wolde													
B8-112r31	on	gewinnē													
B8-112r32	on	suchen													
B8-112r33	loüffin	vnd													
B8-112r34	den	noch													
B8-112r35	mit	eym													
B8-112r36	stet	.													
B8-112r37	du	kanit													
B8-112r38	dorynne	vm													
B8-112r39	den	pfil													
B8-112r40	ylenne	vf													
B8-112r41	so	halt	das	ylen	stete	vnd	heb	on	ein	teil	vber	sich	in	dem	pfile
B8-112r42	So	czihelt	u	on	mit	dem	treue	herulz	vnd	ilt	vil	beller	wä		

-- leer --

Lemma Suche in Grimm Angabe sicher
herau~ ziehen

Part-of-speech Tag
VVFIN: 2, sg., präs., ind., stark

Person Num... Temp... Modus Flexion
2 sg. präs. ind. stark

Anmerkung
 Annotierung noch unbestätigt

Annotierung speichern **Abbrechen**

Abbildung 1: Annotationsdialog für eine Wortform in Lemmano

Gegenüberstellung der vier Varianten als Fließtext mit einer übersichtlichen Form des Apparats.

Ausgangspunkt hier sind digitale Faksimiles, die als Scan der Erstausgaben angefertigt wurden. Diese wurden mittels existierender Software zur Texterkennung in eine maschinenlesbare Form gebracht, anschließend von den Romanisten von fehlerhaft erkannten Stellen bereinigt sowie mit speziellen Markierungen versehen, die beispielsweise Überschriften oder Seitennummern kenntlich machen. Erleichtert wird die Suche nach Fehlern dabei durch ein softwaregestütztes Verfahren, das auffällige Kombinationen von Symbolen anzeigt. So liefert beispielsweise die Suche nach Kombinationen aus Buchstaben und Zahlen ohne trennendes Leerzeichen eine Reihe von fehlerhaft kodierten Jahreszahlen, wie „1506“ oder „1580“. Aus den so entstehenden Textdokumenten wird automatisch eine TEI-konforme XML-Darstellung generiert, die als Grundlage für die folgenden Arbeitsschritte dient. Für einen der beiden algorithmischen Schwerpunkte aus Sicht der Informatik, die Alignierung der Absätze, werden derzeit verschiedene automatische Verfahren geprüft. Für den zweiten Schwerpunkt, die Bestimmung und Visualisierung der Unterschiede auf Absatzebene, wurde bereits ein erster Ansatz implementiert. Dieser vergleicht eine beliebige Anzahl von Textpassagen untereinander. Auf Basis der Levenshtein-Distanz werden die Differenzen zwischen den Varianten ermittelt und daraus eine synoptische Dar-

stellung in LaTeX erzeugt (Abbildung 2). Die Philologen können für die Visuali-

1770

1774

1780

1820

Les ministres¹ de cette princesse² prirent d'abord pour un vifomnaire³ un homme qui voulut⁴ découvrir un monde. Ils le traitèrent longtemps⁵ et cela l'amena à faire⁶ que les hommes communs, quand ils furent en place, ont pour les hommes d'u génie. Colomb ne fut pas rebuté par les difficultés. Il avait⁷ comme tous ceux qui forment des projets extraordinaires¹⁰ cet enthousiasme¹¹ qui les rôdit contre les jugemens de l'ignorance,¹² les dédains de l'orgueil, les petitesfes¹³ de l'avarice, les délaiss de la pareffe,¹⁴ Son ame¹⁵ ferme, élevée, courageuse, fa¹⁶ prudence et son adrefet¹⁷ le firent enfin triompher de tous ces obstacles.¹⁸ On lui accorda trois petits vaiffeaux,¹⁹ et quatre-vingt-dix²⁰ hommes. Il partit le 3 Aout 1492²¹ avec le titre d'Amiral²² et de Vice-roi²³ des îles²⁴ des terres qu'il découvriroit.²⁵

Les ministres¹ de cette princesse² prirent d'abord pour un vifomnaire³ un homme qui voulut⁴ découvrir un monde. Ils le traitèrent longtemps⁵ et cela l'amena à faire⁶ que les hommes en place affectent fi l'avanture, ceux qui n'ont que d'u génie. Colomb ne fut pas rebuté par les difficultés. Il avait⁷ comme tous ceux qui forment des projets extraordinaires¹⁰ cet enthousiasme¹¹ qui les rôdit contre les jugemens de l'ignorance,¹² les dédais de l'orgueil, les petitesfes¹³ de l'avarice, les délaiss de la pareffe.¹⁴ Son ame¹⁵ ferme, élevée, courageuse, fa¹⁶ prudence et son adrefet¹⁷ le firent enfin triompher de tous ces obstacles.¹⁸ On lui accorda trois petits vaiffeaux,¹⁹ et quatre-vingt-dix²⁰ hommes. **S**uite fait le Claude, dont il a moment ne con-²¹
tut pas celle d'una frans, il m²² la veille
le 3 Aout 1492²¹ avec le titre d'amiral²² et de vice-roi²³ des îles²⁴ des terres qu'il découvriroit, et arriva aux Canaries²⁵ où il s'estoit proposé de relâcher.²⁶

Les ministres¹ de cette princesse² prirent d'abord pour un vifomnaire³ un homme qui voulut⁴ découvrir un monde. Ils le traitèrent longtemps⁵ et cela l'amena à faire⁶ que les hommes en place affectent fi l'avanture, ceux qui n'ont que d'u génie. Colomb ne fut pas rebuté par les difficultés. Il avait⁷ comme tous ceux qui forment des projets extraordinaires¹⁰ cet enthousiasme¹¹ qui les rôdit contre les jugemens de l'ignorance,¹² les dédais de l'orgueil, les petitesfes¹³ de l'avarice, les délaiss de la pareffe.¹⁴ Son ame¹⁵ ferme, élevée, courageuse, fa¹⁶ prudence et son adrefet¹⁷ le firent enfin triompher de tous ces obstacles.¹⁸ On lui accorda trois petits vaiffeaux,¹⁹ et quatre-vingt-dix²⁰ hommes. **S**uite fait le Claude, dont il a moment ne con-²¹
tut pas celle d'una frans, il m²² la veille
le 3 Aout 1492²¹ avec le titre d'amiral²² et de vice-roi²³ des îles²⁴ des terres qu'il découvriroit, et arriva aux Canaries²⁵ où il s'estoit proposé de relâcher.²⁶

1770	ministres	1770	princesse	1770	vifomnaire	1770	voulut	1770	trouver long-tem	1770	infante	1770	commun
1 H74	ministres	2 H74	princesse	3 H74	vifomnaire,	4 H74	voulut	5 H74	trouver long-tem	6 H74	infante	7 H74	en place affectent fi l'avanture
H80		H80		H80		H80		H80		H80		H80	
H20	ministres	H20	princesse	H20	vifomnaire	H20	voulut	H20	trouver long-tem	H20	infante	H20	en place affectent fi l'avanture
H70		H70		H70		H70		H70		H70		H70	
8 H74		9 H74	avoit,	10 H74	extraordinaires,	11 H74	enthousiasme	12 H74	Ignorance,	13 H74	pettefes	14 H74	parfe.
H80		H80	avoit,	H80	extraordinaires,	H80	enthousiasme	H80	Ignorance,	H80	pettefes	15 H74	ame
H20		H20	avoit,	H20	extraordinaires,	H20	enthousiasme	H20	Ignorance,	H20	pettefes	H20	ame
H70	fon adrefet	H70	ses obstacles.	H70	sauveur,	H70	quatre-vingt-dix	H70	Il partit	H70	Il partit	H70	commun,
17 H74	fon adrefet,	18 H74	les obstacles.	19 H74	vaiffeaux	20 H74	quatre-vingt-dix	21 H74	Il partit	H80	Sur cette folle escale, dont l'armement ne coitait pas cent mille francs, il mit à la voile	H80	quand il fuit en place, car pour les hommes de
H80	fon adrefet,	H80	les obstacles.	H80	navires	H80	quatre-vingt-dix	H80	Il partit	H80	Sur cette folle escale, dont l'armement ne coitait pas cent mille francs, il mit à la voile	H80	qui n'ont que du
H20	son adrefet,	H20	les obstacles.	H20	navires	H20	quatre-vingt-dix	H20	Il partit	H20	Sur cette folle escale, dont l'armement ne coitait pas cent mille francs, il mit à la voile	H20	qui n'ont que du
H70		H70		H70		H70		H70		H70		H70	
22 H74	Aout 1492,	23 H74	d'amiral	24 H74	d'amiral	25 H74	vise-roi	26 H74	iles et	27 H74	iles et	28 H80	découvert,
H80	Aout 1492,	H80	d'amiral	H80	d'amiral	H80	vise-roi	H80	iles et	H80	découvert,	H80	découvrir,
H20	Aout 1492 ,	H20	d'amiral	H20	vise-roi	H20	vise-roi	H20	iles et	H20	découvrir,	H20	et arriva aux Canaries où il s'étoit proposé de relâcher.

Abbildung 2: Automatisch generierte Synopse mit Variantenapparat für einen Absatz des franz. Textes

sierung der Textvarianten zwischen konfigurierbaren Darstellungsarten wählen. Die genannten Arbeitsschritte, von der Generierung der XML-Dateien bis hin zum Entwurf der elektronischen Edition, sollen perspektivisch in einer gemeinsamen, webbasierten Arbeitsumgebung eingebettet werden.

Prototypische Werkzeuge, unter anderem *Lemmano*, werden in den nächsten Monaten zur Demonstration als Webanwendungen öffentlich verfügbar gemacht.

Anmerkungen

Diese Arbeit wurde durch das Bundesministerium für Bildung und Forschung (BMBF) [Projektkürzel: 01UG1247 / human-325-010 / SaDA] im Rahmen des Projekts „Semi-automatische Differenzanalyse von komplexen Textvarianten“ unter Leitung von Prof. Dr. Thomas Bremer, Prof. Dr. Paul Molitor, Dr. Jörg Ritter und Prof. Dr. Hans-Joachim Solms gefördert. An dieser Stelle möchten wir auch unseren Projektmitarbeiterinnen Sylwia Kösser, Dr. Aletta Leipold und Susanne Schütz danken.

Informatik und Hermeneutik

Erste Erkenntnisse aus dem heureCLÉA-Projekt¹

In unserem aktuellen BMBF-eHumanities Projekt heureCLÉA arbeiten wir als Informatiker und Literaturwissenschaftlerinnen an einer digitalen Heuristik, die die Analyse von literarischen Texten unterstützen soll. Der Anwendungsfokus liegt dabei exemplarisch auf narratologischen Phänomenen der Zeit: d.h., das heuristische Modul von heureCLÉA soll die Funktionalität der Textanalyse und -annotationsumgebung CATMA² erweitern, indem es den Usern automatisch generierte Vorschläge zur Annotation narratologisch definierter Zeit-Phänomene in einem Text anbietet. Das Modul wird auf der Basis von drei Zugängen entwickelt: (1) Ausgangspunkt ist so genanntes "hermeneutisches Markup" (Piez 2010), das auf klassischen narratologischen Kategorien wie etwa Ordnung, Frequenz und Dauer beruht (vgl. Genette 1972, Lahn und Meister 2013) und von geschulten Annotatorinnen vergeben wird. Dieses Markup wird (2) mit regelbasierten Verfahren sowie (3) Machine-Learning-Ansätzen kombiniert.³

Aufgrund des Zusammenspiels von literaturwissenschaftlichen – und speziell: hermeneutischen – Verfahren und informatischen Verfahren der Information Extraction und der Statistik stehen sich non-deterministische Zugänge zu Texten und entscheidbare bzw. deterministische Verfahren gegenüber, die nicht ohne weiteres auf den jeweilig anderen Ansatz übertragen werden können. Deshalb ist die Reproduzierbarkeit von narratologischen Analysen für die Vereinbarkeit des literaturwissenschaftlichen und des informatischen Zugangs und damit für den Erfolg des heuristischen Moduls ausschlaggebend.

In unserem Beitrag präsentieren wir ein methodisches Desiderat im Bereich der Narratologie, das erst durch die interdisziplinäre Zusammenarbeit zwischen Geisteswissenschaftlern und Informatikern in den Fokus gerückt ist und aus unserer Sicht exemplarisch für eine solche Zusammenarbeit ist: die Notwendigkeit, narratologische Analysekategorien eindeutiger zu konzeptionalisieren, um sie operationalisieren zu können.

¹ vgl. www.heureclea.de (gesehen am 10.12.2013)

² vgl. www.catma.de (gesehen am 10.12.2013)

³ Damit ist heureCLÉA ein Beitrag zur *computational narratology* im Sinne von Mani, da es zu "exploration and testing of literary hypotheses through mining of narrative structure from corpora" (Mani, 2013, para. 1) beiträgt.

³Zu den regelbasierten Verfahren vgl. Strötgen und Gertz (2010), die Gesamtarchitektur von heureCLÉA wird außerdem in einem weiteren eingereichten Beitrag vorgestellt.

Die Narratologie ist eine literaturwissenschaftliche Disziplin, die eine Reihe theoretischer Konzepte und Modelle für die Analyse und Interpretation erzählender Texte zur Verfügung stellt. Diese narratologischen Kategorien dienen normalerweise der Bezeichnung und Verortung textueller Eigenschaften, die (a) als typisch für narrative Texte angesehen werden und (b) für besonders interessant und geeignet befunden werden, um die speziellen Eigenschaften eines literarischen Einzelwerkes herauszustellen. Viele der Kategorien dienen der Bezeichnung struktureller Phänomene, die hauptsächlich an der Textoberfläche zugänglich sind. Das gilt insbesondere auch für die meisten Kategorien, die der Analyse explizit markierter Zeitphänomene dienen, wie sie im Rahmen von *heureCLÉA* untersucht werden.⁴ Obwohl narratologische Kategorien gemeinhin als theoretisch durchdacht und leicht operationalisierbar gelten, zeigten sich bei ihrer formalisierten Anwendung im Rahmen manueller, kollaborativer Annotation in *heureCLÉA* einige theoretische Unzulänglichkeiten. Typischerweise wurden solche Unzulänglichkeiten dann entdeckt, wenn sich die Annotatoren hinsichtlich der korrekten narratologischen Bestimmung konkreter Textstellen nicht einig waren. In Diskussionen über die Gründe für individuelle Annotations-Entscheidungen stellte sich dann oft die uneindeutige oder unvollständige Konzeption der jeweiligen Kategorie als Ursache uneinheitlichen Markups heraus. Die festgestellten theoretischen Versäumnisse lassen sich in zwei Gruppen einteilen, die je unterschiedliche Problemlösungsstrategien erfordern:

a) konzeptionelle Unvollständigkeit, die leicht durch eine Vervollständigung der Kategorie mittels funktionaler Entscheidungen behoben werden kann. Stellt sich bei der versuchten Anwendung einer Kategorie heraus, dass ihre Definition zu vage ist, um die Bestimmung einer fraglichen Textstelle vorzunehmen, müssen pragmatische Entscheidungen im Hinblick auf die Inklusion oder Exklusion bisher nicht bedachter textueller Oberflächenmerkmale getroffen werden.⁵

⁴ Der Klarheit wegen sollte angemerkt werden, dass keine der im Feld der Narratologie interessanten Phänomene rein formale Textmerkmale sind, da die *Bedeutung* von Wörtern und Sätzen stets ausschlaggebend für ihre Bestimmung ist. Das bedeutet, dass solche Phänomene zwar an der Textoberfläche zugänglich sind, ihre Bestimmung jedoch trotzdem in einem weiteren Sinne des Wortes interpretativ sein kann.

⁵ Ein derartiger Problemfall stellte ich an folgender Textstelle in Friedrich Hebbels Erzählung *Matteo* im Hinblick auf die Frage, ob es sich hier um eine Prolepse - bisher konzeptionalisiert als Vorgriff in der Zeit - handelt: "Sieh, morgen feire ich meine Hochzeit; zum Zeichen, daß du mir nicht mehr böse bist, kommst du auch, meine Mutter wird dich gern sehen." (Hebbel 1963: para. 4). Die Schwierigkeit ist hier dadurch gegeben, dass die angesprochene Figur am folgenden Tag nicht auf der Hochzeit erscheint. Um

Derartige Entscheidungen haben nur für die Anwendung der jeweiligen Kategorie Konsequenzen, nicht aber für weitere Konzepte. - Die zweite Kategorie betrifft dagegen

b) theoretische Unvollständigkeit, die ihrerseits auf die unzureichende Bestimmung fundamentaler narratologischer Konzepte zurückzuführen ist. Probleme dieses Typs können nicht einfach durch pragmatische Entscheidungen behoben werden, weil die für eine Problemlösung notwendigen Setzungen auf der Ebene grundlegender narratologischer Konzepte weitreichende Konsequenzen für viele erzähltheoretische Einzelkategorien nach sich zieht. Im Folgenden soll diese zweite Problemklasse anhand eines Beispiels erläutert werden.

In der Erzählung *Der Tod* von Thomas Mann ist bei dem Vergleich der Annotationsergebnisse im Hinblick auf die Geschwindigkeit der Erzählung⁶ folgende Passage in den Fokus der Aufmerksamkeit gerückt:

Ich habe die ganze Nacht hinausgeblickt, und mich dünkte, so müsse der Tod sein oder das Nach dem Tode: dort drüben und draußen ein unendliches, dumpf brausendes Dunkel. Wird dort ein Gedanke, eine Ahnung von mir fortleben und -weben und ewig auf das unbegreifliche Brausen horchen? Mann 2004: 76

Diese Passage wurde von einigen Annotatoren ab dem ersten Komma als zeitraffend erzählt eingeordnet, von anderen dagegen als Erzählpause. Die Diskussion über die Gründe für die individuellen Entscheidungen hat gezeigt, dass die Annotatoren unterschiedliche Auffassungen darüber vertreten, was ein Ereignis ist. Betrachtet man mentale Vorgänge als Ereignisse, so muss man die zitierte Passage als zeitraffend klassifizieren, da die Gedanken des Erzählers in der fiktiven Welt vermutlich längere Zeit anhielten als die wenigen Sekunden, die in der Erzählung für ihre Wiedergabe eingeräumt werden. Ist man jedoch der Ansicht, dass es sich bei mentalen Prozessen nicht um Ereignisse handelt, so liegt in obiger Textstelle eine Pause vor: Der Bericht von Ereignissen wird unterbrochen durch die Darstellung nicht-ereignishafter Gegebenheiten. Die Frage danach, welche Konzeption von Ereignis korrekt oder sinnvoll ist, ist Gegenstand der Debatte um Narrativität: die für erzählende Texte konstitutive Eigenschaft, von Ereignissen zu berichten. Die unterschiedlichen Intuitionen der Annotatoren in Bezug auf die Definition von "Ereignis" korreliert hier mit Schmids Konzeptionen von Ereignis I, das jegliche

entscheiden zu können, ob hier eine Prolepsis vorliegt, muss entschieden werden, ob dieses Konzept auch antizipierte Ereignisse fassen soll, die im Verlauf der Erzählung nicht eintreten.

⁶ Unter "Erzählgeschwindigkeit" versteht man in der Narratologie das Verhältnis zwischen der Menge an Ereignissen, von denen berichtet wird, und der Zeit, die für diesen Bericht notwendig ist.

Form von Zustandsveränderung inkludiert, und Ereignis II, das zusätzliche Kriterien anführt, die Zustandsveränderungen aufweisen müssen, um als Ereignis zu gelten (Schmid 2003).⁷ Eine Entscheidung im Hinblick auf die richtige Narrativitätsdefinition, die für die Lösung von Annotationsproblemen im Bereich der Erzählgeschwindigkeit notwendig wäre, hätte nun nicht nur für die fraglichen Kategorien Konsequenzen, sondern beispielsweise auch für die Bestimmung des Gegenstandsbereich der Narratologie und potenziell für eine Reihe weiterer Kategorien.⁸

Angesichts insbesondere dieser zweiten Sorte von Problem stellt sich die Frage, inwieweit solche grundlegenden Fragen im Rahmen von *heureCLÉA* geklärt werden können und sollten. Da die theoretische Arbeit an narratologischen Grundkonzepten nicht im Fokus des Projektes stehen sollte, war zunächst ein individueller Umgang der Annotatoren mit den anwendungsbezogenen Einzelproblemen vorgesehen. Es hat sich jedoch herausgestellt, dass diese Vorgehensweise weder aus narratologischer Sicht befriedigend ist, noch eine aus informationstheoretischer Perspektive verwertbare Datengrundlage liefert. Aus diesen Gründen haben wir uns dazu entschieden, den beiden geschilderten narratologischen Basisproblemen einige Aufmerksamkeit zu widmen: Für die Bestimmung von Ebenenwechsel und -zuordnung wird eine konsistente Lösung gefunden, so dass Ordnungsphänomene tatsächlich unterschiedlichen Erzählebenen zugeordnet werden können. Für die Bestimmung von „Ereignis“ streben wir eine plausible Konzeptionalisierung an, die ein möglichst wenig interpretatives Erkennen von Ereignissen erlaubt.

⁷ Zu diesen Kriterien zählt neben Resultativität, Relevanz, Unvorhersehbarkeit, Effekt, Irreversibilität und Nicht-Wiederholbarkeit auch das Kriterium der Faktizität, das die Eigenschaft von Zustandsveränderungen bezeichnet, tatsächlich in der fiktiven Außenwelt stattzufinden. Wertet man Faktizität als notwendige Eigenschaft von Ereignissen, muss die oben zitierte Passage aus *Der Tod* als Erzählpause klassifiziert werden.

⁸ Eine ähnliche Verknüpfung von Annotationsproblemen und ungeklärten narratologischen Basiskonzepten findet sich bei der Annotation von Phänomenen der zeitlichen Ordnung einer Erzählung einerseits und dem grundlegenden narratologischen Konzept der Erzählebenen. Es kann nur sinnvoll das zeitliche Verhältnis von solchen Ereignissen bestimmt werden, die sich auf derselben Erzählebene befinden. Anhand welcher Faktoren ein Ebenenwechsel festzumachen ist, wird in der narratologischen Forschung noch diskutiert (vgl. Ryan 1991, Coste/Pier 2011).

Die beschriebene Problematik ist ein spezifisch literaturwissenschaftliches Problem, die sie erzeugenden Rahmenbedingungen sind jedoch zugleich exemplarisch für das Zusammenspiel von Informatik und Geisteswissenschaften. Deshalb ist der entwickelte Lösungsansatz von entscheidender Bedeutung für das Gelingen des Projekts. Die Reproduzierbarkeit von Analyseergebnissen, die durch den Ansatz anvisiert wird, wird in den Geisteswissenschaften traditionell nicht thematisiert, da diese meist dem Konzept der intersubjektiven Übereinstimmung operieren, ohne diese weiter zu bestimmen. Die Reproduzierbarkeit von Analyseergebnissen ist jedoch zugleich auch eine von mehreren, bislang wenig erforschten Gelingensbedingungen für interdisziplinäre Projekte im Bereich der *Digital Humanities*. Diese disziplinäre Doppelperspektive auf ein methodisches Kriterium weist insofern auf die konzeptionellen Chancen, Probleme und Bedingungen einer Kooperation zwischen Geisteswissenschaftlern und Informatikern im Kontext von DH-Projekten.

Literatur

Coste, D. and Pier, J. (2013). Narrative Levels. *the living handbook of narratology*.
<http://www.lhn.uni-hamburg.de/article/narrative-levels> (gesehen am 06.12.2013). First published 2011.

Genette, G. (1972). Discours du récit. In id., *Figures III*. Paris: Editions Du Seuil, pp. 67-282.

Lahn, S. and Meister, J. C. (2013). *Einführung in die Erzähltextranalyse: 2nd, updated edition*. Stuttgart: Metzler.

Mani, I. (2013). Computational Narratology. *the living handbook of narratology*.
<http://www.lhn.uni-hamburg.de/article/computational-narratology> (gesehen am 06.12.2013).

Piez, W. (2010): Towards Hermeneutic Markup: an Architectural Outline. *Digital Humanities 2010. Conference Abstracts*. London: Office for Humanities Communication, Centre for Computing in the Humanities, King's College London, pp. 202-205.

Ryan, M.-L. (1991). *Possible Worlds, Artificial Intelligence, and Narrative Theory*. Bloomington: Indiana UP.

Schmid, W. (2003): Narrativity and Eventfulness. In T. Kindt & H.-H. Müller (eds.). *What Is Narratology? Questions and Answers Regarding the Status of a Theory*. Berlin: de Gruyter, 17–33.

Strötgen, J. and Gertz, M. (2010). HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. *Proceedings of the 5th International Workshop on Semantic Evaluation (ACL 2010)*. Uppsala, pp. 321-324.

KI und Geisteswissenschaften

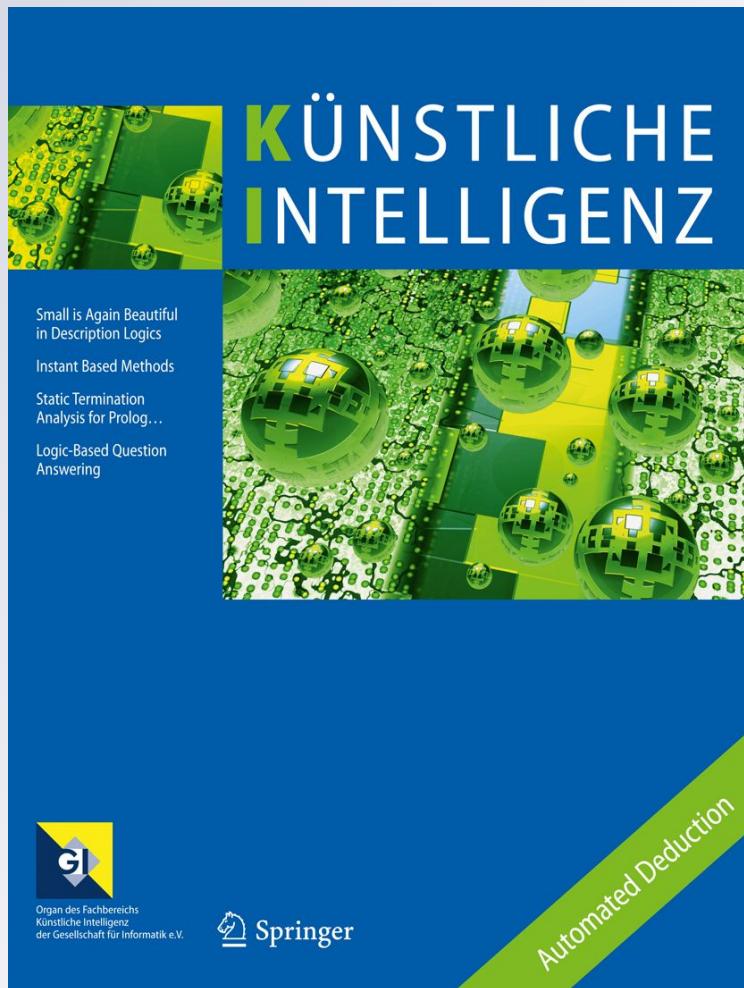
Günther Görz

KI - Künstliche Intelligenz

Organ des Fachbereichs "Künstliche Intelligenz" der Gesellschaft für Informatik e.V. - German Journal on Artificial Intelligence

ISSN 0933-1875
Volume 25
Number 4

Künstl. Intell. (2011) 25:313–315
DOI 10.1007/s13218-011-0128-5



 Springer

Your article is protected by copyright and all rights are held exclusively by Springer-Verlag. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your work, please use the accepted author's version for posting to your own website or your institution's repository. You may further deposit the accepted author's version on a funder's repository at a funder's request, provided it is not made publicly available until 12 months after publication.

KI und Geisteswissenschaften

Günther Görz

Eingegangen: 25. Juli 2011 / Angenommen: 22. September 2011 / Online publiziert: 4. Oktober 2011
© Springer-Verlag 2011

Die Beziehungen zwischen „Künstlicher Intelligenz“ und den Geisteswissenschaften sind vielfältiger, als es auf den ersten Blick erscheinen mag, wenn man an KI als Teilgebiet der Informatik, also zunächst als technisches Fach, denkt. Dass die KI aus der technischen Perspektive Besonderheiten aufweist, die über die Informatik im Allgemeinen hinausweisen, kann mit Recht bezweifelt werden; auch die von Wolfgang Wahlster hin und wieder vorgetragene Lesart von „KI“ als „künftiger Informatik“ scheint dem Recht zu geben. In historischer Perspektive gibt es eine Tradition, spätestens mit Leibniz beginnend, über Babbage, Turing und Zuse – um nur einige der Informatik-Pioniere zu nennen –, in der in den ersten Schritten zur Informatik bereits programmatische Elemente der KI zu finden sind. Jede kritische Reflexion der KI wäre damit im Prinzip nichts anderes als eine kritische Reflexion der Informatik, und wir wären mit unseren Überlegungen bereits ans Ende gekommen: Zur KI als avantgardistischer Vorhut der Informatik, wie sie sich zuweilen gerne sieht, getreu der Devise „If it works, it’s no longer AI“, könnte nur festgehalten werden, dass sie manche Grundannahmen etwas pronaconierter und in einigen ihrer Vertreter wie Marvin Minsky oder Hans Moravec auch radikaler formuliert, aber nichts grundsätzlich Neues einbringt... Wäre da nicht ihr fundamental interdisziplinärer Ansatz, auf den in den einschlägigen Lehrbüchern immer wieder hartnäckig verwiesen wird, und der doch über das rein Technische und den Szentismus Minskyscher Provenienz hinauszugehen scheint.

Diesen weiteren Horizont bedenkend sei also noch einmal gefragt: Gibt es Wechselwirkungen zwischen KI und

den Geisteswissenschaften? Wohlgemerkt, Wechselwirkungen, womit Einflüsse in beiden Richtungen gemeint sind. Es ist klar, dass an dieser Stelle kein Definitionsversuch der „Geisteswissenschaften“ erfolgen kann, sondern wir einfach von dem durchaus heterogenen Fächerkanon ausgehen, der an den philosophischen Fakultäten unserer Universitäten trotz aller neoliberalen Demontageversuche immer noch vorhanden ist. Gemeinsam ist ihnen ihr Selbstverständnis als Orientierungswissenschaften: Es geht um das Verstehen von Handlungen, um Erklären, Rekonstruieren und Planen.

Auch eine Eingrenzung der KI als wissenschaftliche Disziplin ist nicht einfach, und eine fundierte wissenschafts-historische Auseinandersetzung damit wäre absolut wünschenswert, doch wollen wir uns für das Folgende mit der breit angelegten denk- und handlungsbezogenen Charakterisierung, wie sie sich im Lehrbuch von Russell und Norvig findet, zufrieden geben. Die fundamentale Abstraktionsleistung der KI könnte man damit in der Schematisierung und Automatisierung von Verstandesleistungen sehen. Angesichts der intellektualistischen Operationalisierungen, die die KI seit den 1950er Jahren entwickelte, wurde immer intensiver eingefordert, auch auf Emotionalität und Verkörperung einzugehen, doch trotz zahlreicher Bemühungen ist hier nach wie vor erhebliche Grundlagenarbeit zu leisten. An dieser Stelle ist ersichtlich auf weitgehende Überschneidungen mit der Kognitionswissenschaft und ihre psychologischen Wurzeln hinzuweisen.

Die Wechselwirkungen zwischen KI und Geisteswissenschaften sollen im Folgenden in zwei Bereichen etwas genauer in den Blick genommen werden: Zum einen geht es um einige Hinweise zur Fundierung der sich als interdisziplinär verstehenden KI, hauptsächlich aber um einschlägige Anwendungen, insbesondere im Rahmen der „Digital Humanities“ und des Semantischen Web, das hoffentlich trotz aller Kommerzialisierung auch die Chance für ein Epistemi-

G. Görz (✉)
Friedrich-Alexander-Universität Erlangen-Nürnberg,
91058 Erlangen, Deutschland
e-mail: goerz@cs.fau.de

sches Web mit freiem Zugang für alle zum Wissen dieser Welt bieten wird.

Schon früh gab es eine fundamentale Kritik an den Prämissen und Forschungsansätzen der KI aus der Philosophie, wobei insbesondere das Buch von Dreyfus einen hohen Welenschlag verursacht hat. So einseitig und verengt die vorgebrachte Kritik aus der Warte von Phänomenologie und Hermeneutik auch gewesen sein mag: Die wütende pauschale Ablehnung aus der KI-Gemeinde und damit die Verweigerung einer sachlichen Auseinandersetzung lag wohl vor allem an durchaus nachvollziehbaren förderpolitischen Gründen. Was wäre aus der KI geworden, wenn die militärischen Freunde von ARPA eingesehen hätten, dass der Forschungsansatz der KI schneller an seine Grenzen stösst als gedacht und sie deshalb die Gelder in andere Kanäle lenkten? Zumindest die reflektierte Neuorientierung von Terry Winograd hätte doch zu bedenken geben sollen, dass mit formalen Mitteln allein ein adäquates Handlungsverständen nicht zu erreichen sein wird. Neben der konzeptionellen Kritik gab es noch eine Reihe anderer Einwände, vor allem politischer Art, die jedoch wenig KI-spezifisch sind, sondern auf die Informatik als Ganzes zutreffen – was angesichts der o.g. Gemeinsamkeiten in der frühen Entwicklung des Fachs auch nicht erstaunlich ist. Auf die Fundamentalkritik an der KI in grundsätzlicher Weise einzugehen würde bedeuten, eine zufriedenstellende Antwort auf das psychophysische Problem geben zu können, und das sieht doch sehr nach einer endlosen Geschichte aus.

Angesichts der für die KI relevanten kognitiven Leistungen wie Wahrnehmung, Gedächtnis, Analyse, und Synthese wird es für die Kognitionswissenschaft, sofern sie sich nicht nur naturwissenschaftlich-reduktionistisch versteht – also Handlungsverständen im Sinne der Hermeneutik thematisiert –, problematisch mit Ansätzen zur Operationalisierung: Wenn man Handeln auf Verhalten reduziert, ist man gleich wieder bei den Naturwissenschaften. Naturalistische Hermeneutik ist eine *contradictio in adjecto*. Die ganze Problematik wird ja deutlich an der Diskussion um die Rolle bzw. den Erkenntnisgewinn bildgebender Verfahren: Sie zeigen nur Korrelationen, aber keine Kausalität — auch wenn Letzteres von wissenschaftstheoretisch wenig gebildeten Vertretern der Zunft immer wieder frech, aber medienwirksam behauptet wird.

Stattdessen möchte ich vorschlagen, eine fruchtbringende Synthese in einer Epistemologisierung kognitionswissenschaftlicher Ansätze zu suchen, d.h. in einer Reinterpretation auf der Wissensebene: Welche Wissensarten sind denn im Spiel, wenn bestimmte kognitionswissenschaftliche Modelle konstruiert werden, und von welchem Wissenstyp sind diese Modelle selbst? Das ginge in die Richtung einer epistemischen Anreicherung des „Linguistic Turn“ unter Berücksichtigung der historischen Dimension. Nelson Goodman hat das einmal so ausgedrückt, dass die Struktur des

Geistes – also der Mentalismus – abgelöst wird durch die Struktur der Begriffe und diese wiederum durch die verschiedenen Symbolsysteme der Wissenschaften und der Philosophie. Damit sich nicht alles in postmoderner Beliebigkeit auflöst, ist auf die Einheit der wissenschaftlichen Rationalität zu verweisen. Diese wird praktisch garantiert durch die Verfahren der Verifikation und der bedeutungskonstituierenden Begründung in der (Wissenschafts-) Sprache.

Was ergibt sich aus der Fundierungsdebatte für das Verhältnis von KI und Geisteswissenschaften? Eine echte Wechselwirkung gibt es nur in geringem Umfang, aber die KI kann lernen, anstelle der Verteidigung universeller Erklärungsansprüche in einen rationalen Diskurs über Wissensformen einzutreten. Und dabei hat sie aus langer Beschäftigung mit Wissensrepräsentation und -verarbeitung auch durchaus methodisch etwas zu bieten, z.B. kann sie bei der Bereitstellung eines sprachlichen Rahmens zur begrifflichen Modellierung und Begründung helfen und Hilfsmittel zur Klassifikation, Annotation oder Verknüpfung von Wissenselementen beitragen.

Damit ist eine Brücke zum Anwendungsbereich geschlagen: „Digital Humanities“ – als „Wissenschaften vom Verstehen“, wie jüngst der Untertitel einer Fachtagung lautete – ist eine immer häufiger benutzte Sammelbezeichnung für Informatikanwendungen in den Geistes- und Kulturwissenschaften – wohl wissend, dass das angelsächsische Verständnis von „Humanities“ nicht mit dem der „Geisteswissenschaften“ deckungsgleich ist.

Mit der von formalen Mitteln geprägten Erforschung symbolischer Strukturen, wie sie uns in Antike, Mittelalter, früher Neuzeit und in außereuropäischen Kulturen begegnen, ist weitgehend Neuland zu betreten. Begriffssysteme haben ihre ihnen eigene historische Dynamik: Es geht um die Entwicklung und den Einsatz von Standards für Namen, Bezeichnungen, Fachtermini und ihre Systematisierung in Thesauri und formalen Ontologien, die zugleich ihren – durch vielfältige Transferbeziehungen beeinflussten – historischen Wandel mitbedenken. Generell richtet sich an die KI die Frage, welche Beiträge ihre Forschungsansätze und Methoden, etwa zur Wissensrepräsentation und Inferenz, zur informierten Suche, zur Textanalyse und semantischen Erschließung, zur Bildanalyse und Objekterkennung, zur Visualisierung komplexer Datenstrukturen, zur Planung und Problemlösung sowie zu kognitiven Aspekten liefern können.

„Digital Humanities“ unter Einschluss semantischer Techniken stünde dann in einem Verhältnis zu den Geisteswissenschaften wie „Computational Science“ zu den Naturwissenschaften. Es wäre verfehlt, darin die Aufkunft einer neuen Wissenschaft zu sehen – man bedenke nur, was aus den Ansprüchen der vermeintlichen Superwissenschaft Kybernetik aus den 1950er Jahren geworden ist. Dass andererseits die KI eine reine Hilfswissenschaft bleiben muss, ist noch lange nicht gesagt. Lehrreich mag hier

sein, auf das Verhältnis von Computerlinguistik und der KI-nahen Sprachverarbeitung („Natural Language Processing“) zu schauen: Die Sprachverarbeitung, die in den „Digital Humanities“ selbst zahlreiche Anwendungen findet, hat durchaus eine eigene Akzentuierung. Sie unterscheidet sich von der Computerlinguistik vor allem im Erkenntnisinteresse und in der Anwendung, auch wenn sich beide der gleichen algorithmischen Mittel bedienen: Die Sprachverarbeitung hat kein primäres Interesse an grammatischen Strukturen aus linguistischer Sicht, sondern sieht sie als eine Zwischenrepräsentation auf dem Weg zur semantisch-pragmatischen Inhaltsanalyse – mitgeteilte Bedeutungen in gesellschaftlichen Kommunikationsprozessen, die in symbolischer Form repräsentiert werden. Das schließt den Einsatz stochastischer Methoden nicht aus, aber letztlich geht es nicht um Wahrscheinlichkeitsverteilungen, sondern um in formalsprachlicher Form repräsentierte kommunikative Inhalte.

Darüber hinaus kann die KI vielfältige Angebote bei der Modellbildung und Simulation auf der Basis der Erfahrung über mehrere Jahrzehnte mit deskriptiven und algorithmisch-prozeduralen Beschreibungsverfahren beisteuern. Die Wissensverarbeitung – Repräsentation und Inferenz – ist in besonderer Weise im Kontext des Semantischen Web wieder aufgenommen worden und hat durch die dort vorgenommene Standardisierung auf der Basis von XML durchaus für zahlreiche Korpusprojekte, um nur diese zu nennen, an Attraktivität gewonnen. Plötzlich eröffnen sich kompatible Anschluss- und Erweiterungsmöglichkeiten, die auch durch die neu hinzugewonnenen Datensets

beachtliche Härtetest für KI-Verfahren versprechen. Auch was unser Verständnis von Inferenz betrifft, wird sich einiges ändern: Neben die Explikation impliziten Wissens tritt die Verarbeitung von Massendaten v.a. mit statistischen Verfahren, was mit deren Verfügbarkeit durch das Internet und grid-förmig organisierter massiver Verarbeitungskapazität erst seit der Mitte der neunziger Jahre möglich geworden ist. So wird mittelfristig der Einsatz der digitalen Techniken wohl auch zu einer Veränderung der Forschungsmethoden und -strategien in den jeweiligen geisteswissenschaftlichen Disziplinen führen, so dass hier eine echte Wechselwirkung zu erwarten ist.

Mit dem Gewicht auf der Semantik, auf Systematik und Methodologie zeichnet sich eine Entwicklung von einfachen, teilweise uninformativen Verweisstrukturen zu Argumentationszusammenhängen ab und damit besteht die Chance zu einem Epistemischen Web, das diesen Namen verdient. Allerdings darf dabei die pragmatische Dimension nicht aus dem Blickfeld geraten; die kontinuierliche Konsensbildung zwischen den wissenschaftlichen Fachgemeinschaften und Kulturen ist notwendig — auch dies lehrt die Auseinandersetzung mit der kulturellen Überlieferung.

Danksagung Der Autor dankt Georg Hohmann, Siegfried Krause, Stefan Mandl und Martin Scholz für hilfreiche Hinweise.

Günther Görz ist Professor für Informatik (KI) an der Friedrich-Alexander-Universität Erlangen-Nürnberg, Zweitmitglied der Philosophischen Fakultät und Gastwissenschaftler am Max-Planck-Institut für Wissenschaftsgeschichte, Berlin.

Wittgensteins Nachlass: Computerlinguistik und Philosophie

Der Finder wiTTFind und die Wittgenstein Advanced Search Tools (WAST)

Max Hadersbeck, Alois Pichler, Florian Fink, Øyvind Liland Gjesdal

Maximilian.Hadersbeck@lmu.de

Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München,

Wittgenstein Archives at the University of Bergen (WAB).

1 EINLEITUNG

In meinem Vortrag möchte ich über unsere Arbeitsgruppe „Wittgenstein in Co-Text“ am Centrum für Informations- und Sprachverarbeitung (CIS) der Ludwig Maximilians Universität München (Dr. Max Hadersbeck) und dem Wittgenstein-Archiv an der Universität Bergen (WAB) / Norwegen (Dr. Alois Pichler) berichten. Vor zwei Jahren begannen wir in dieser Arbeitsgruppe für die öffentlich zugänglichen Teile des Nachlasses von Ludwig Wittgenstein (siehe Bergen Electronic Edition (BEE, 2000) und die Open Source Plattform *Wittgenstein Source* (<http://www.wittgensteinsource.org>, 2009-)) computerlinguistische Verfahren zu entwickeln, die einen neuen WEB-basierten Zugang zu den Texten ermöglichen, um Wörter und Phrasen im „Zusammenhang des Satzes“ zu finden. Denn so schrieb schon Wittgenstein im *Tractatus logico-philosophicus* (3.3): „Nur der Satz hat Sinn; nur im Zusammenhang des Satzes hat ein Name Bedeutung“.

WAB und CIS entwickelten in enger Zusammenarbeit ein einfaches TEI-P5 konformes XML-Format (CISWAB), das einen optimalen Ausgangspunkt für die Zusammenarbeit von Wittgensteinforschern und Computerlinguisten darstellt.

Dazu extrahierten wir aus CISLEX, dem am CIS erstellten Vollformenlexikon des Deutschen, das Speziallexikon wiTTLex. In intensiven Gesprächen mit Computerlinguisten, Editionsspezialisten, Informatikern, Philosophen, einem interdisziplinären Seminar und einer zweitägigen Sommerschule („Digital Wittgenstein Scholarship 2013“) in München wurden der wissenschaftliche Austausch gepflegt und nach und nach Fragestellungen der Wittgensteinforscher in computerlinguistische Verfahren umgesetzt. Diese Verfahren fassten wir unter der Bezeichnung „Wittgenstein Advanced Search Tools“ (WAST) zusammen und implementierten sie in unserem WEB-basierten Finder WiTTFind.

In meinem Vortrag möchte ich WiTTFind vorstellen, die zugrunde liegenden Wittgenstein Advanced Search Tools beschreiben und über unsere Erfahrungen mit der Kooperation Computerlinguistik und Philosophie berichten.

In der folgenden Abbildung zeigen wir das Eingabefeld unseres Finders WiTTFind:

Siehe <http://wittfind.cis.uni-muenchen.de>:

The screenshot shows the WiTTFind search interface. At the top, there is a search bar with the text "WiTTFind sagen". To the right of the search bar is a button labeled "WiTTFind-Suche". Below the search bar, there is a link "WiTTFind-Suche". The main content area displays a snippet from Wittgenstein's Tractatus. The snippet includes a header with links to "Faksimile", "Wittgenstein Source Normalized", and "Wittgenstein Pundit". The text of the snippet reads: "[7] 6) Man sagt: ein Wort verstehen heißt, wissen, wie es gebraucht wird. Was heißt es, das zu wissen? Dieses Wissen haben wir sozusagen im Vorrat. (S. 22)". Below this, a larger block of text is shown in a greenish background, with the first line being "6) Man sagt: ein Wort verstehen heißt, wissen, wie es gebraucht wird." and the second line being "Was heißt es, das zu wissen? Dieses Wissen haben wir sozusagen im Vorrat. (S. 22)". The word "B E D E U T U N G !" is printed below the text in a stylized font.

2 ÖFFENTLICH ZUGÄNGLICHE TEXTE DES NACHLASSES

2.1 TEXT: DAS TEI-P5 KONFORME XML-FORMAT (CISWAB)

Die am WAB in Bergen entstandene XML-Transkription des Nachlasses von Ludwig Wittgenstein annotiert die Texte sehr detailliert: Alle Streichungen, Ergänzungen usw. sind im XML festgehalten. Diese genaue Auszeichnung ist aber für den Einsatz unseres Finders viel zu ausführlich, und so definierten wir ein reduziertes TEI-P5 konformes XML-Format (CISWAB), das eine geeignete Basis für die Zusammenarbeit von Wittgensteinforschern und Computerlinguisten darstellt. CISWAB wird über XSLT-Transformation aus dem umfassenderen WAB XML extrahiert (Øyvind Liland Gjeddal, WAB).

2.2 LEXIKON: DAS ELEKTRONISCHE VOLLFORMENLEXIKON MIT SEMANTISCHEN WORTKLASSEN (WITTLex)

Das Vollformenlexikon WiTTLex umfasst alle Wörter der auf Wittgenstein Source öffentlich und frei zugänglichen Texte des Nachlasses von Ludwig Wittgenstein und ist im DELA Format, das am Laboratoire d'Automatique Documentaire et Linguistique (LADL, Paris) definiert wurde, gespeichert. Für jedes Wort werden im Lexikon die Vollform, das Lemma, die lexikographische Wortform, semantische Notationen und morphologische Varianten gespeichert (Angela Krey, CIS).

3 DIE WEB-BASIERTE APPLIKATION ZUM FINDEN VON TEXTSTELLEN: WITTFIND

Im Zentrum unserer computerlinguistischen Arbeit entwickelten wir hocheffiziente, parallelisierte C++ Client/Server-Programme (Florian Fink, CIS), die die XML-notierten Texte einlesen, das Vollformenlexikon im Hintergrund halten und alle WAST-Verfahren implementiert haben. Über eine WEB-Schnittstelle können Anfragen gestellt werden, und das Finder-Programm sucht regelbasiert nach Textstellen, die zu dieser Anfrage passen. WEB-Programme bereiten die Ergebnisse für die HTML-Ausgabe auf, extrahieren die zugehörigen Faksimileausschnitte und stellen die Treffer auf der WEB-Seite dar.

4 SUCHANFRAGEN BEI WITTFIND

4.1 LEMMATISIERTE UND INVERSE LEMMATISIERTE WORT-SUCHE

Mit Hilfe der Einträge im Vollformenlexikon WiTTLex können Anfragen an WiTTFind lemmatisiert behandelt werden. Die Suche nach dem Wort „sagen“ liefert z.B. alle Textstellen, an denen lexikalische Varianten wie „sagte“, „sagten“ usw. vorkommen. Das Lexikon erlaubt auch eine inverse lemmatisierte Suche: Die Anfrage „sagte“ führt zum Lemma „sagen“, und daraus werden alle lexikalischen Varianten von „sagen“ generiert und danach gesucht.

4.2 SUCHE ÜBER WORTFORMEN

Im Vollformenlexikon WiTTLex sind alle Wörter mit ihrer Wortform gespeichert. Nomen und Adjektive sind semantisch annotiert. Unser Finder erlaubt es, nach Wörtern zu suchen, die diesen Wortformen zugehören. Die Anfrage: „Die <ADJ> Farbe“ findet zum Beispiel alle Sätze, welche die Wortfolge „Die“ gefolgt von einem Adjektiv und dem Wort „Farbe“ enthalten. Die Anfrage: „die <COL>“ findet alle Sätze mit der Wortfolge „die“-gefolgt-von-einer-Farbe, und als weiteres Beispiel findet die Anfrage mit semantischer Wortform: „die <EN>“ alle Sätze, welche die Wortfolge „die“-gefolgt-von-einem-Eigenamen (Olga Strutynska, CIS) enthalten.

4.3 SATZSTRUKTUR UND WILDCARDS

Diese Suche erlaubt dem Nutzer festzulegen, dass ein bestimmtes Wort am Satzanfang <BOS>, bzw. Satzende <EOS> vorkommt. Stellvertretend für ein Wort oder eine Zeichenkette kann auch der Wildcard-Buchstabe „*“ verwendet werden. Die Anfrage: „<BOS> Ich den“ findet alle Sätze, die mit „Ich“ beginnen und von Wörtern gefolgt werden, die mit „den“ beginnen. Die Anfrage: „<BOS> Ich * * * <EOS>“ findet zum Beispiel alle Sätze, die mit „Ich“ beginnen und vier Wörter lang sind.

4.4 REGELBASIERTE LINGUISTISCHE SUCHE UND PART OF SPEECH TAGGING (POS-TAGGING)

Hier wurde die Möglichkeit der deutschen Sprache implementiert, dass bei Partikelverben die Partikel vom Wortstamm getrennt vorkommen können. Partikelverben werden über unser Lexikon WiTTLex erkannt und zerlegt (Luidmilla Volos, CIS). Um Partikel von Präpositionen zu disambiguiieren, verwendet unser Finder ein automatisches Part of Speech Tagging (treetagger von Dr. Helmut Schmid, CIS) und lokale Grammatiken. Die lokalen Grammatiken wurden mit dem graphischen WEB-Tool CisGraph (Shuangjiao Cao, Medieninformatik), der ebenfalls von uns

programmiert wurde, erstellt. Zum Beispiel findet die Anfrage „nachdenken“ jetzt auch Sätze wie: „Wir denken nie darüber nach, ...“

4.5 SUCHE OHNE ALTERNATIVEN

Ein charakteristisches Merkmal von Wittgensteins Nachlass besteht darin, dass er in den Texten sehr viel änderte und oftmals mehrere alternative Formulierungen anbietet. In den vorliegenden Texten existieren also viele alternative Lesarten, die am WAB in XML kodiert sind. Damit die Suche auch in allen unterschiedlichen Lesarten durchgeführt werden kann, werden bei unserer Suche im Hintergrund alle Lesarten der Texte generiert, durchsucht, und die gefundenen Textstellen mit ihren Alternativen dargestellt (Patrick Seebauer, CIS).

5 DARSTELLUNG DER GEFUNDENEN TEXTSTELLEN AUCH IM FAKSIMILE

Gerade bei sehr heterogenen Schriftensammlungen, wie die des Nachlasses von Ludwig Wittgenstein, bei der viele, auch handschriftliche, Texte, vorliegen und vom Autor häufig geändert wurden, ist es für die Wissenschaftler sehr wichtig, die gefundenen Textstellen nicht nur als edierten Text zu sehen, sondern die zugehörigen Stellen auch im Faksimile der Originale studieren zu können. Nur im Bild des Originals bekommen die Forscher die „Aura“ des gefundenen Textes zu spüren, und mit Hilfe des Faksimiles können sie sogar Editionsfehler entdecken. Am CIS (Matthias Lindiger) wurde ein Programm entwickelt, welches die Textedition und das Faksimile auf Bemerkungenniveau miteinander verlinkt, und welches es daher erlaubt, von einem Suchergebnis direkt zum entsprechenden Ausschnitt im Faksimile zu springen.

6 VERBINDUNG ZU BESTEHENDEN SOFTWARETOOLS AUS DER WITTGENSTEINFORSCHUNG

Es war uns von Anfang an wichtig, dass unser Tool keine neue digitale Insellösung darstellen sollte, sondern dass eine einfache Verbindung zu bestehenden Plattformen und Programmen herstellbar sein soll. Jede gefundene Textstelle kann durch Anklicken sofort in Wittgenstein Source dargestellt werden und mit dem Semantic Web tool Pundit (<http://feed.thepund.it>) weiter annotiert werden.

7 ZUSAMMENARBEIT COMPUTERLINGUISTIK UND PHILOSOPHIE

Die Zusammenarbeit zwischen dem CIS, dem WAB und Wittgensteinforschern (u.a. an der Fakultät Philosophie LMU, Department II) ist sehr intensiv und für beide Seiten sehr anregend.

Die Computerlinguisten realisierten, dass Texte, nicht wie bei herkömmlichen Suchmaschinen, statistische Ereignisse sind, sondern dass bei dieser Art von Texten jedes Wort wichtig ist und die Trefferquote bei 100% liegen muss: „Nichts darf übersehen werden!“. Außerdem erkannten die Computerlinguisten, dass nur solche Tools und Verfahren von den Kooperationspartnern akzeptiert werden, die einen einfachen WEB-basierten Zugang ermöglichen, sich an der Wissenschaftssprache der Wittgensteinforscher orientiert und für deren dahinterliegenden Formalisierungen offen oder sogar frei konfigurierbar sind. Das war ein Grund, weshalb wir von Anfang an einen regelbasierten Ansatz für unseren Finder wählten.

Die Kooperationspartner der Philosophie schätzen die technische Möglichkeit, dass die gefundenen Textstellen in den Faksimiles der Originale, die meist unzugänglich in Archiven aufbewahrt sind, sichtbar gemacht werden können. Nichtdeutschsprachige Forscher sind von der Lemmatisierung ihrer Anfrage sehr erfreut, da sie oftmals der Reichhaltigkeit der deutschen Morphologie nicht in dem Maße mächtig sind. Die Möglichkeit der gleichzeitigen Recherche über mehrere Dokumente hinweg erlaubt es hervorragend, Ähnlichkeiten und Textgenesen zu entdecken, was schon immer ein zentraler Forschungsgegenstand der Wittgensteinforschung war.

So wird aus der Zusammenarbeit der Computerlinguisten und der Wittgensteinforscher ein ständiger Kreislauf, bei dem die Philosophen zur Formalisierung ihrer Fragestellungen aufgefordert werden, die Computerlinguisten Verfahren entwickeln müssen, welche die anspruchsvollen Formalisierungen effizient implementieren, und die neuen Anfragemöglichkeiten wiederum zu einer erneuten Korrektur und Erweiterung der Formalisierung des Findens führen.

Wie schreibt schon Ludwig Wittgenstein im Ms111,178: "Wenn ich etwas suche, so ist es wesentlich, daß ich das Finden ebenso ausführlich muß beschreiben können (ob es (je so) eintritt oder nicht) ehe der Gegenstand gefunden ist.“

So nennen wir unser Tool WiTTFind nicht eine Suchmaschine, sondern einen Finder.

Pre-Conference am 25./26. März 2014

„DARIAH-DE – Aufbau von Forschungsinfrastrukturen für die e-Humanities“

im Rahmen der

DHd-Konferenz am 26. – 28. März 2014

„DH – methodischer Brückenschlag oder 'feindliche Übernahme'? Chancen und Risiken der Begegnung zwischen Geisteswissenschaften und Informatik“

Universität Passau

Die von Dariah-DE organisierte zweitägige Pre-Conference verfolgt das Ziel zwei zentrale Themenkomplexe anhand der inhaltlichen Säulen der Dariah-DE Forschungsinfrastruktur mit Hilfe von Workshop-Sessions zu thematisieren. Hierbei stehen folgende Leitfragen im Mittelpunkt:

1. Welche methodischen, thematischen und technologischen Anforderungen haben forschende Geistes- und KulturwissenschaftlerInnen an digitale Forschungsinfrastrukturen und welche Bedeutung haben diese Anforderungen in der Lehre?
2. Wie können digitale Forschungsinfrastrukturen nachhaltig jenseits befristeter Projektförderzeiträume institutionell etabliert werden?

Für Dariah-DE besteht eine digitale Forschungsinfrastruktur für die Geistes- und Kulturwissenschaften aus vier Komponenten: Lehre, Forschung, Forschungsdaten und technische Infrastruktur. Eine bloße Reduzierung auf technische Aspekte würde verhindern, dass die Anforderungen und Bedürfnisse von WissenschaftlerInnen aus den Geistes- und Kulturwissenschaften beim Aufbau einer digitalen Forschungsinfrastruktur beachtet würden. Aus diesem Grund sollen aktuelle Entwicklungen und Perspektiven jenseits eines Projektberichts thematisiert und mit den TeilnehmerInnen in einzelnen Workshop-Sessions diskutiert werden, mit dem Ziel forschungsbezogene und -relevante Anforderungen aus geistes- und kulturwissenschaftlicher Perspektive zu benennen.

Neben vier inhaltlichen Sessions zu den Kernelementen der Dariah-DE Forschungsinfrastruktur – Lehre, Forschung, Forschungsdaten und technischer Infrastruktur –, wird in einer abschließenden fünften Session der Umgang mit Objektdaten, wie sie z.B. in der Archäologie und anderen bild- und objektanalysierenden Kulturwissenschaften als Forschungsgegenstand verwendet werden, thematisiert und die daraus resultierenden Herausforderungen für digitale Forschungsinfrastrukturen, die sich überwiegend bislang auf Texte und Quellen fokussierte, analysiert.

Parallel hierzu ist geplant, dass an beiden Tagen das Dariah-DE Café stattfindet, bei dem die TeilnehmerInnen der Pre-Conference die Möglichkeit haben, einzelne Projekte und Forschungsvorhaben, die Komponenten der Dariah-DE Forschungsinfrastruktur nutzen und weiterentwickeln, sich präsentieren zu lassen. Vor allem der Dialog mit EntwicklerInnen, WissenschaftlerInnen und VertreterInnen von Forschungsinfrastrukturen steht hierbei im Mittelpunkt und soll die Möglichkeit des technischen, inhaltlichen, methodischen und interdisziplinären Austauschs geben.¹ Hierdurch erhalten die TeilnehmerInnen die Chance, mit inhaltlich verwandten Projekten zu diskutieren und – auch neue Kontakte aufzubauen. Die Präsentationen der mit Dariah-DE assoziierten Forschungsprojekte, die aus unterschiedlichen fachwissenschaftlichen Disziplinen und institutionelle Kontexten (Universitäten, Akademien, außeruniversitäre Forschungseinrichtungen etc.) stammen, werden entweder als Poster, als Live-Präsentationen oder als Demo-Sessions erfolgen. Darüber hinaus ist geplant, dass in diesem Rahmen auch die europäische Verortung zu Dariah-EU und anderen internationalen Forschungsvorhaben vorgestellt werden. Zugleich werden von Dariah-DE-VertreterInnen entwickelte fachwissenschaftliche Dienste, wie z.B. der Geo-Browser und die Collection Registry, weitere Komponenten der technischen Infrastruktur und curriculare Themen, in Planung befindliche DH-Studiengänge und forschungsbezogene Ergebnisse präsentiert. Studentische Gruppen von verschiedenen Universitäten werden darüber hinaus eigene Forschungsprojekte und ihre aktuellen Arbeiten vorstellen.

Als Abschluss des ersten Tages findet ein Abendvortrag von Dr. Karl-Heinz Mörth, Institute for Corpus Linguistics and Text Technology, Austrian Academy of Science, statt. Hierbei werden die derzeitigen Entwicklungen beim Aufbau von digitalen Forschungsinfrastrukturen für die Geistes- und Kulturwissenschaften in Österreich und deren Verbindung und Kooperationen zu den gemeinsamen Entwicklungen innerhalb des europäischen Forschungsraums thematisiert.

Kontaktdaten

Dariah-DE – Digitale Forschungsinfrastrukturen für die Geisteswissenschaften
Dr. Heike Neuroth
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Papendiek 14
37073 Göttingen
neuroth@sub.uni-goettingen.de

Zahl der erwarteten Teilnehmer

Ca. 50 Personen

Benötigte Technische Ausstattung

- Vorträgsraum für die Workshop-Sessions
- 2. Raum für Dariah-DE Café
- Technische Standardausstattung für Vorträge: Beamer, Leinwand etc.
- ca. 8 Monitore (23"+) für Präsentationen
- Stellwände für das Dariah-DE Café

¹ Eine Auswahl an Projekten, die anhand von Demo-Sessions oder mit Postern ihre aktuellen Entwicklungstätigkeiten vorstellen und bereits angefragt wurden bzw. noch angefragt werden sollen, findet sich im Anhang. Weitere Projekte sollen im Laufe des Januars noch angefragt werden.

**„DARIAH-DE –
Aufbau von Forschungsinfrastrukturen für die e-Humanities“**

Programm

Dienstag, 25. März 2014

<i>Uhrzeit</i>	<i>Thema</i>
13:00-13:30	Begrüßung und Eröffnung Dr. Heike Neuroth (SUB Göttingen)
13:30-15:00	1. Session Forschungsfragen und -methoden Dr. Christof Schöch (Uni Würzburg), Dirk Wintergrün (MPIWG Berlin)
15.00-15:30	Kaffeepause
15.30-16.30	2. Session Lehre Prof. Dr. Manfred Thaller (Uni Köln)
16.30-18.00	Demo-Session Teil I (in einem gesonderten Raum) - Demo-Sessions im Rahmen des DARIAH-DE-Cafés - Demo-Sessions assoziierter DARIAH-DE Projekte
ab 18.00	Abendvortrag Dr. Karlheinz Mörtl, Österreichische Akademie der Wissenschaften „Der Aufbau von digitalen Forschungsinfrastrukturen für die Geistes- und Kulturwissenschaften in Österreich“
ab 20.00	Gemeinsames Conference Dinner

Mittwoch, 26. März 2014

<i>Uhrzeit</i>	<i>Thema</i>
9:00-9:15	Eröffnung des 2. Tages
9:15-10:30	3. Session Wissenschaftliche Sammlungen Dr. Thomas Stäcker (HAB Wolfenbüttel)
10.30-10.45	Kaffeepause
10.45-12.00	4. Session Technische Infrastruktur Peter Gietz (DAASI), Tibor Kalman (GWDG, Göttingen)
12.00-12.30	5. Session Objekt-Cluster Prof. Dr. Reinhard Förtsch (DAI Berlin)

12.30-13.00	Abschlussdiskussion: „Digitales Forschen und Lehren in den Geisteswissenschaften – Themen und Perspektiven für Dariah-DE“ Dr. Heike Neuroth (SUB Göttingen)
13.00-14.00	Demo-Sessions Teil II (in einem Raum) <ul style="list-style-type: none"> - Demo-Sessions im Rahmen des Dariah-DE-Cafés - Demo-Sessions assoziierter Dariah-DE Projekte

Dariah-DE Café: Assoziierte Forschungsprojekte (Auswahl)

- Dariah-EU
- Relationen im Raum – Visualisierung topographischer Klein(st)strukturen (RiR)
- eCodicology – Algorithmen zum automatischen Tagging mittelalterlicher Handschriften
- ePoetics – Korpuserschließung und Visualisierung deutschsprachiger Poetiken (1770-1960) für den Algorithmic criticism
- NELi – Vernetzte Korrespondenz. Erforschung und Visualisierung sozialer, räumlicher, zeitlicher und thematischer Netze in Briefkorpora
- 3D-Joins und Schriftmetrologie
- SlaVaComp – COMPutergestützte Untersuchung von Variabilität im KirchenSLAvischen
- epidat – epigraphische Datenbank / digitales Textarchiv
- Freischütz Digital
- Beethovens Werkstatt – Genetische Textkritik und digitale Edition
- Das sächsisch-magdeburgische Recht als kulturelles Bindeglied zwischen den Rechtsordnungen Ost- und Mitteleuropas
- Fontane Notizbücher
- Blumenbach Online
- TextGrid / Shared Canvas
- DH-Nachwuchsgruppenprojekt „Computergestützte literarische Gattungsstilistik“
- Studentische Projekte aus Bamberg, Würzburg, Darmstadt, Köln
- Edition des rabbinischen Auslegungsmidrash zu den Psalmen, Midrash Tehillim
- Personendatenrepositorium

Using Ontologies as Heuristic Tools: Sources in the History of Philosophy and Their Interpretation in the Semantic Web

In my presentation, I will first criticise the ongoing debate on whether folksonomies or ontologies are more powerful tools for the organisation and representation of knowledge. In this, I rely on a basic insight from the philosophy of language regarding the mutual dependence of the meaning of concepts and the meaning of propositions. Based on this criticism, I want to argue for a new methodology of building ontologies ‘inductively’. I will finally apply this methodology to a use case, namely the semantic representation of sources in the history of early modern philosophy.

In the debate between ‘ontologists’ and ‘folksonomists’, ‘ontologists’ base their view mostly on the conviction that the world as it is is ‘classifiable’. Their ‘folksonomist’ opponents have some reservations, mostly because they operate in domains in which the practical applicability of ontological schemata is at best dubious. I contend that this debate is based on a fundamentally flawed semantic assumption: a ‘building block’ account of the meaning of propositions as a mere sum total of the meaning of their parts. This overemphasis on the ‘compositionality’ of propositional meaning leads to a neglect of the complementary aspect of ‘contextuality’, i.e. the insight that the meaning of concepts as the constituent parts of propositions is at least to a certain extent determined by the very propositions they appear in. This means that nonpropositional uses of concepts and singular terms are parasitical on the use of concepts and singular terms in propositions.

Hence, an act of folksonomist ‘tagging’ contains at best the assertion of an incomplete proposition, so that, from the point of view of the philosopher of language, ‘tags’ are systematically underdetermined, because in themselves such labels cannot clarify the exact relation between the tagged entity and the label used to describe or characterise it. Conversely, ontologists cannot presume that the meaning of the concepts they employ is fully stable over uses of these concepts in different propositional contexts. This is particularly true as soon as we assess the practical value of semantic web technologies in the context of the digital humanities.

Prime examples for the practical success of the top-down approach favoured by the ‘ontologists’ can be found in the sciences, based on their clear commitment to the existence of natural kinds. But as soon as we enter the sphere of the cultural, the surplus of such a methodological approach is more questionable, the impressive achievements of ontologies like CIDOC CRM notwithstanding. The practical value of such ontologies for the digital humanist is limited, since the complexity of the domain is mirrored in a similarly complex conceptual schema that may be accessible only to ‘classification experts’ like librarians or cultural heritage

professionals, experts that often share the conviction of the natural scientist that their respective domain is 'classifiable in principle'.

But although both the ontologist and the folksonomist argue from faulty premises, there is no need to abandon all attempts to represent knowledge about the cultural in machine-readable form. These insights should rather prompt us to revise our common strategies of ontology building for cultural artifacts. Changes to how we devise representations of conceptual structures would allow us to respect the semantic interdependence of concepts and propositions, to use ontologies as heuristic tools for the exploration of an unknown conceptual space, and to combine the explicitness and precision of ontologies with the flexibility of folksonomies. My talk discusses the basic features of such a methodology under the heading of 'inductive ontology building'.

For this, it is first helpful to contemplate again the question why we need (or want) an ontology of a given domain. We can use an ontology as an instrument for structuring information about a domain or we can apply it for automated reasoning tasks. However, in both cases, the basic unit of information within an ontology is the proposition, so that the primary class of entities in an inductive ontology are propositions and the facts they express. So we first need a basic and sparse 'ontology of propositions' in order to represent these particulars. Such an ontology would know only one class, namely propositions. It would contain three relations, namely 'has-subject-term', 'has-relation-term', and 'has-object-term'. In this model concepts in a proposition are not to be interpreted as atomic building blocks, but as dependent aspects of the proposition they are used in. Therefore, a proposition in this model should be expressed not in one, but in three RDF triples:

- 'S:Proposition-p R: 'has-subject-term' O:literal
- 'S:Proposition-p R: 'has-relation-term' O:literal
- 'S:Proposition-p R: 'has-object-term' O:literal

Such a schema respects the basic syntactical requirement of RDF that subject terms and relation terms in RDF statements must be resources, i. e. identifiable via a URI. At the same time, such a form of representation need not make any assumptions whatsoever about the conceptual structure of the underlying domain.

We can then begin to identify e. g. terms within a given domain that are predominantly employed in the subject position. We can extract facts in which different subject terms appear with identical relation or object terms, inferring either identical extensions or class hierarchies. By identifying 'outliers' standing in no relation to other terms, we can circumscribe the domain in question more clearly and ask whether facts that do not share terms with any other collected fact really belong to the domain under examination.

An example may help to clarify the advantages of such a strategy. At <http://emto-nanopub.referata.com>, I have begun to extract doxographical content from early modern sources, notating these excerpts as triples. In my presentation, I will talk about 62 propositions that were asserted by 16th and 17th century Spanish philosophers while debating the proper definition of philosophy and relate them to the vocabulary available in the ontology language OWL for describing the relations between the concepts employed. OWL offers resources to describe and characterise the relation between classes, properties, subclasses, subproperties, equivalent properties, or instances. An ontology of how 17th century Spanish philosophers conceptualised philosophy would start from the assertion that philosophy is a habit, i.e. as an acquired property of the soul, so that we could tentatively ‘philosophy’ as an OWL ‘subproperty’ of OWL ‘property’ ‘habit’, ‘habit’ as a subproperty of the property ‘quality’, and ‘quality’ as a property of the class ‘soul’. Philosophical disciplines could again be introduced as subproperties of the property ‘philosophy’. I will then show how to model the relation between a proposition and an author entertaining this proposition and how this model can be used to make ontologies author-relative.

In closing, I will briefly compare this ‘digital’ way of reading sources to traditional approaches of reconstructing historical sources using the complete apparatus of first-order predicate logic. The main advantage of the methodology proposed here is a reduction of complexity. We are not concerned with the logical reconstruction of arguments, but only with the collection and analysis of statements. This reduction of complexity is particularly relevant when charting unknown territory. Since mass digitisation has made available vast amounts of completely unknown sources in the last decade, the history of early modern philosophy faces enormous challenges in coming to terms with this material. In the light of current research on OCR technologies for early modern prints it can be expected that a significant percentage of these sources will soon be available as full text, so that we may be able to develop tools for an automatic extraction of RDF triples describing their content. The methodology proposed here may then prove to be a fruitful strategy for turning this content into semantically rich information.

Der Weg des Texts zum Nutzer - Historische Quellen der Jüdischen Studien im semantischen Web

Rachel Heuberger

Weltweit nehmen die digitalen Ressourcen für Jüdische Studien zu. Das Konsortium Judaica Europeana hat bislang einige Millionen digitaler Objekte erfasst, die das jüdische Leben in Europa dokumentieren und diese in die Europeana importiert. Viele der digitalen Objekte wurden von der Judaica Sammlung der Universitätsbibliothek Frankfurt am Main geliefert, die zu den weltweit bedeutendsten Sammlungen ihrer Art zählt. Die historischen Judaica Bestände der UB Frankfurt wurden digitalisiert, mit Werken aus anderen Bibliotheken ergänzt und werden im Portal der Digitalen Sammlungen Judaica online bereitgestellt.

Das Portal stellt einen digitalen Quellenkorpus dar, der durch Interdisziplinarität und Vielsprachigkeit der präsentierten Medien aus einem Zeitraum von über acht Jahrhunderten gekennzeichnet ist und eine wichtige Infrastruktur für die virtuelle Forschungsumgebung und für internationale Kooperationsprojekte anbietet. Das Portal der Digitalen Sammlungen Judaica, das einheitlich durchsuchbar ist, ist in sieben unterschiedliche Sammlungen, je nach Projektphasen, Finanzierung und Inhalt gegliedert. Die Sammlungen decken verschiedene thematische sowie mediale Aspekte des Forschungsschwerpunktes ab, sind in unterschiedlicher Tiefe und Methodik erschlossen und werden teilweise noch fortgeführt.

Es handelt sich um die Sammlungen:

- Compact Memory, das die 110 wichtigsten jüdischen Zeitungen und Zeitschriften des deutschsprachigen Raumes aus den Jahren 1806-1938 umfasst. Die Periodika repräsentieren die gesamte religiöse, politische, soziale, literarische oder wissenschaftliche Bandbreite der jüdischen Gemeinschaft und stellen für die Erforschung des Judentums in der Neuzeit eine unverzichtbare Quelle dar. Die Zeitschriften wurden aus unterschiedlichen Bibliotheksbeständen digitalisiert, 81.000 Einzelbeiträge von mehr als 10.000 Autoren sind bibliographisch erschlossen.
- Die Freimann-Sammlung, eine virtuelle Rekonstruktion der in aller Welt verstreuten Werke der Vorkriegssammlung der Wissenschaft des Judentums, die in Kooperation mit deutschen Instituten begonnen und gegenwärtig in Kooperation mit dem Center for Jewish History/ Leo Baeck Institute in New York fortgesetzt wird.
- Der Korpus der Quellen in hebräischen Schriftzeichen, getrennt gegliedert in hebräische Handschriften, die die 360 hebräischen Handschriften der Bibliothek

umfassen, hebräische Inkunabeln sowie jiddische Drucke, eine Sammlung von rund 800 historischen Büchern in jiddischer Sprache.

- Die Sammlung Judaica Frankfurt, die sowohl Werke in hebräischen als auch in lateinischen Schriftzeichen umfasst und inhaltlich eine Bandbreite von Werken zur Geschichte der Juden in Frankfurt, Frankfurter hebräische Drucke, hebräische religiöse Schriften des 16. und 17. Jahrhunderts, jiddische (Theater-)Literatur sowie eine Notensammlung jiddischer Lieder abdeckt.
- die Rothschild-Sammlung, ein historisches Unikat von rund 20.000 Zeitungsausschnitten der nationalen und internationalen Presse aus den Jahren 1885-1928 mit inhaltlichen Bezug zur Familie Rothschild in lateinischen Lettern und in europäischen Sprachen.

Insgesamt sind bislang rund 2 Mill. digitalisierter Seiten online abrufbar.

Einzelne Sammlungen, wie die Rothschild-Sammlung und Teile von Compact Memory, sind OCR texterfasst und damit automatisiert durchsuchbar, andere werden manuell strukturiert und intellektuell erfasst. Allen gemeinsam ist die Erschließung mit Metadaten nach internationalen Standards in Formaten, die den Export aus dem Bibliothekssystem in das Datenmodell EDM (European Data Model) der Europeana ermöglichen.

Einzelne Komponenten der Metadaten sind bereits im RDF Format und so Teil der Linked Open Data im semantischen web, allen voran die Personenanzahlung. Durch die Verknüpfung der Personennamen mit der Gemeinsamen Normdatei (GND) der Deutschen Nationalbibliothek, die wiederum in VIAF (Virtual International Authority File) eingebunden ist, wird ein eindeutiges Bezugssystem für die bibliographischen Daten sichergestellt und gleichzeitig eine Anbindung an die Erschließungssysteme anderer Institutionen sowie an das sich ausweitende Datennetz im LOD Format ermöglicht. In Bearbeitung sind zur Zeit weitere Komponenten, die mittels aufbereiteter Vokabulare und Thesauri auch zu einer inhaltlichen Erschließung in RDF führen.

Der Vortrag beabsichtigt die neue digitale Ressource "Digitale Sammlungen Judaica" der UB Frankfurt in ihrer Einbindung in die Europeana, unter besonderer Berücksichtigung der bereits erreichten Erschließungstechniken, darzustellen und die bereits bestehenden vielfältigen Rechercheoptionen innerhalb des Portals sowie in der Verknüpfung mit anderen Portalen aufzuzeigen. Außerdem sollen Möglichkeiten des Einsatzes von weiteren digital gestützten Verfahren ausgelotet werden.

Brauchen die Digital Humanities eine eigene Methodologie?

Überlegungen zur systematischen Nutzung von Text Mining Verfahren in einem politikwissenschaftlichen Projekt

Die Verfügbarkeit neuer Verfahren und Technologien in der Informatik hat in den letzten Jahren weltweit neue Forschungsansätze in den Geistes- und Sozialwissenschaften ermöglicht, die es in einer bisher nicht bekannten Weise erlauben, umfangreiche digitale oder digitalisierte Datenbestände in die geisteswissenschaftliche Forschung einzubeziehen. In der Praxis findet sich jedoch nach wie vor der Fall, dass ein Projektvorhaben entweder eine Anwendung von generischen Verfahren aus der Informatik auf geistes- oder sozialwissenschaftliche Daten darstellt, bei der die Interessen der Fachwissenschaften oft nur ungenügend berücksichtigt werden. Oder das Vorhaben stellt eine von den Fachwissenschaften „dominierte“ Adaption etablierter, nicht-computergestützter Arbeitsweisen nach den Vorgaben der Fachwissenschaftler dar, bei der die Informatik in die Rolle eines bloßen Dienstleisters gedrängt wird. In diesem Zusammenhang ist zunächst die Unterscheidung von *Datenmanagement* und *Datenauswertung* in den Digital Humanities hilfreich. Beschränken sich interdisziplinäre Kooperationen lediglich auf die Digitalisierung, Speicherung und Verwaltung von Daten, sind die Möglichkeiten für eigene Forschungsbeiträge auf Seiten der Informatik eher beschränkt. Einen wesentlichen Beitrag für die Fachwissenschaften können die Digital Humanities aber vor allem dann leisten, wenn digitale Daten nicht nur per Computer verwaltet, sondern auch ausgewertet werden sollen. In diesem Fall stellen sich besondere Anforderungen an die Kooperation, die eigene Forschungsleistungen und tiefere Verständnisse für die Arbeit des jeweils Anderen von den beteiligten Disziplinen verlangen.

Im Rahmen des gemeinsam von Politikwissenschaftlern und Informatikern betriebenen und vom BMBF geförderten Projektvorhabens „ePol - Postdemokratie und Neoliberalismus. Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949-2011“ (Wiedemann/Lemke/Niekler 2013) versuchen wir deshalb, eine Verständigungsebene zu entwickeln, die eine Klärung gegenseitiger Anforderungen und spezifisch auf diese angepasste Lösungen ermöglicht, mit denen Forschungsinteressen auf beiden Seiten angegangen werden können, und die in diesem Sinne Nachhaltigkeit erzeugt, als dass beide Seiten ihre jeweiligen Arbeitsmethoden verändern, anstatt nur mit den Ergebnissen des anderen Teilprojekts umzugehen. Eine solche Verständigungsebene muss aus unserer Perspektive vor allem Aspekte *transdisziplinärer Modellierung* umfassen – mit anderen Worten, Sozialwissenschaftler und Informatiker müssen eine gemeinsame (Definitions-)Sprache finden. Für die Informatik ist die Modellierung essentiell. Denn nur durch die Beschreibung von Sachverhalten in Form theoretischer, prozessualer und mathematischer Modelle, lassen sich Problemlösungen in Form von Computerprogrammen implementieren. Modelle sind aber immer Vereinfachungen der Realität und bedürfen der Anpassung bzw. Korrektur

¹ Abteilung Automatische Sprachverarbeitung, Institut für Informatik an der Universität Leipzig

durch die Fachwissenschaftler. Dies macht die Modellierung zum zentralen Angelpunkt zwischen den Projektbeteiligten.

Im Fall des Projektvorhabens ePol soll die These der Postdemokratie – Politik wird unter dem Einfluss des Neoliberalismus zunehmend von ökonomischen Gesichtspunkten bestimmt - erstmals umfassend empirisch analysiert werden. Grundlage bilden mehr ca. 3,5 Millionen Texte aus Tageszeitungen und Wochenzeitschriften (Zeit, FAZ, taz, SZ) im Zeitraum von 1949 bis 2011, die mit Verfahren der automatischen Sprachverarbeitung ausgewertet werden sollen.

Für die Operationalisierung der politikwissenschaftlichen Fragestellung, und damit der Bereitstellung von Kriterien für die Konzipierung eines Forschungsdesigns und die Auswahl passender Textanalyseverfahren, bieten sich zwei Zugänge an. Zum Einen können wir auf Verfahren in den Sozialwissenschaften zur Hypothesenbildung und deren empirischer Überprüfung aus dem Bereich der Inhaltsanalysen zurückgreifen. Verschiedene Methoden der qualitativen und quantitativen Textanalyse² haben sich in der Politikwissenschaft fest etabliert. Eine Anpassung dieser *Vorgehensmodelle* für (semi-)automatische Analysen großer Textmengen (z.B. vollständige Jahrgänge retrodigitalisierter Tageszeitungen) steht jedoch noch aus. Zum anderen können wir uns an Verfahren des Requirements Engineering aus der Informatik orientieren. Diese Verfahren sind zwar insbesondere in der Wirtschaftsinformatik primär auf die Betrachtung von betriebswirtschaftlich relevanten Prozessen hin orientiert (Kastens et. al. 2008), konnten jedoch auch für die Modellierung von digitalen Kulturgütern (z.B. Texte, Bilder, Musik, Objekte, Filme, Karten) als Gegenstand geistes- und sozialwissenschaftlicher Untersuchungen und für die Bewahrung kulturellen Erbes weiterentwickelt werden (vgl. Schlieder/Wullinger 2010).

Im Rahmen unseres Vortrags skizzieren wir, wie mit Hilfe eines vereinfachten Requirements Engineering die komplexen inhaltlichen Fragen des Projekts ePol in enger Absprache zwischen den Projektbeteiligten in einzelne Teilaufgaben aufgegliedert werden konnten, für die klare funktionale Anforderungen erarbeitet und deren Implementierung mit konkreten Programmpaketen spezifiziert werden konnte (vgl. Abb. 1 und 2).

Die Operationalisierung der Forschungsfrage von Seiten der politischen Theorie erfolgt über die Beobachtung der Veränderungen öffentlichen Begründens von Politikmaßnahmen in Richtung einer zunehmenden „Ökonomisierung“ (Lemke 2012). Um diese Veränderung in medialer Berichterstattung messbar zu machen, wurden im Rahmen der Anforderungsanalyse drei wesentliche Teilaufgaben spezifiziert:

1. Retrieval relevanter Dokumente: Erstellung eines Subkorpus aus den 3,5 Millionen Dokumenten unseres Gesamtkorpus nach den Kriterien einer hohen Dichte neoliberaler Sprachgebrauchsmuster und argumentativer Begründungen,
2. Identifizierung von Argumenten: Manuelle Annotation von Argumenten in unserem Subkorpus; Entwicklung und Anwendung eines Klassifikationsverfahrens, dass weitere (Kandidaten für) Argumente in den Texten automatisch findet,
3. Analyse: Validierung / Falsifizierung von Hypothesen auf den (semi-)automatisch extrahierten Daten; Erstellung geeigneter Visualisierungen zur Analyse der Zeitreihendaten.

2 Etwa die wissenssoziologische Diskursanalyse nach Keller (2007); Zur Unterscheidung von hypothesenprüfender und rekonstruktiver Sozialforschung siehe Bohnsack (2010).

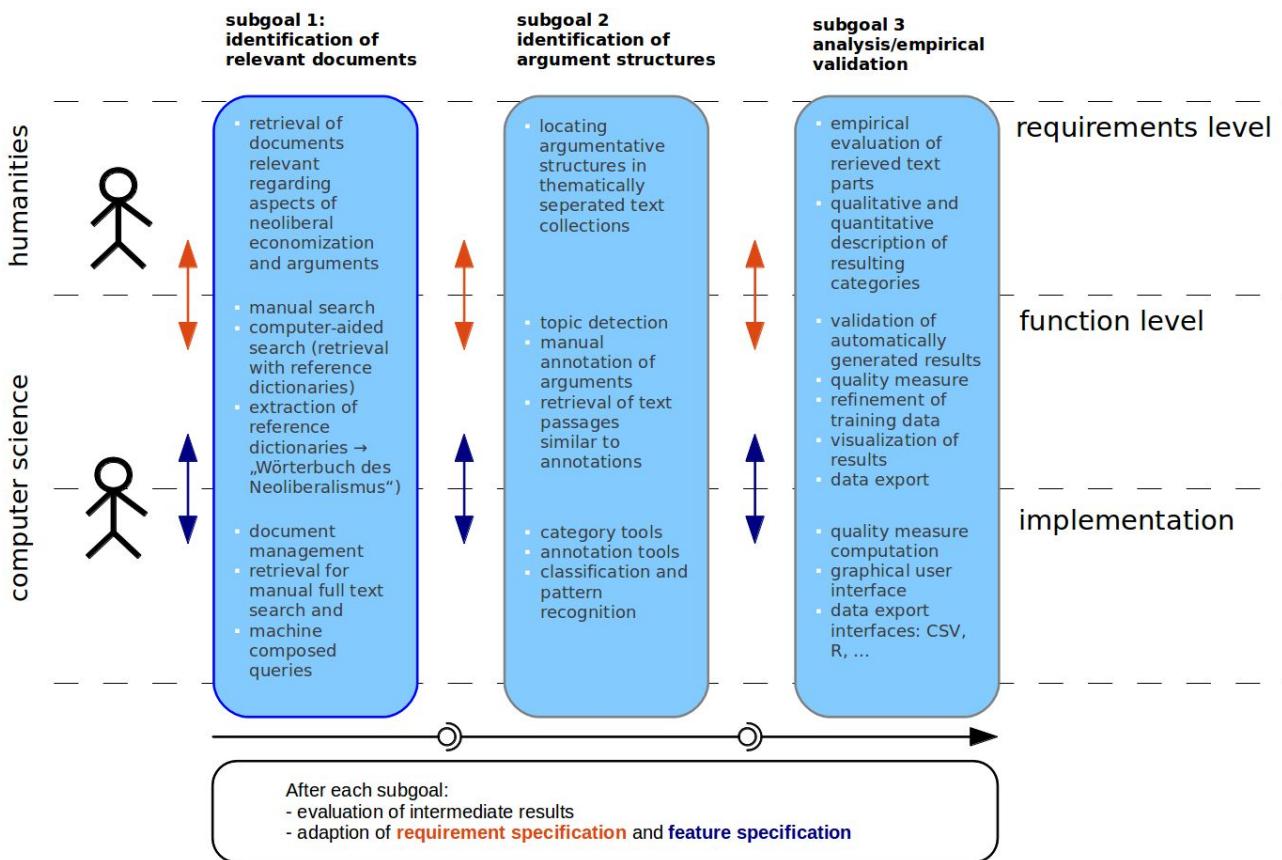


Abbildung 1 – Unterteilung der Anforderungen in Teilaufgaben und Verantwortungsebenen

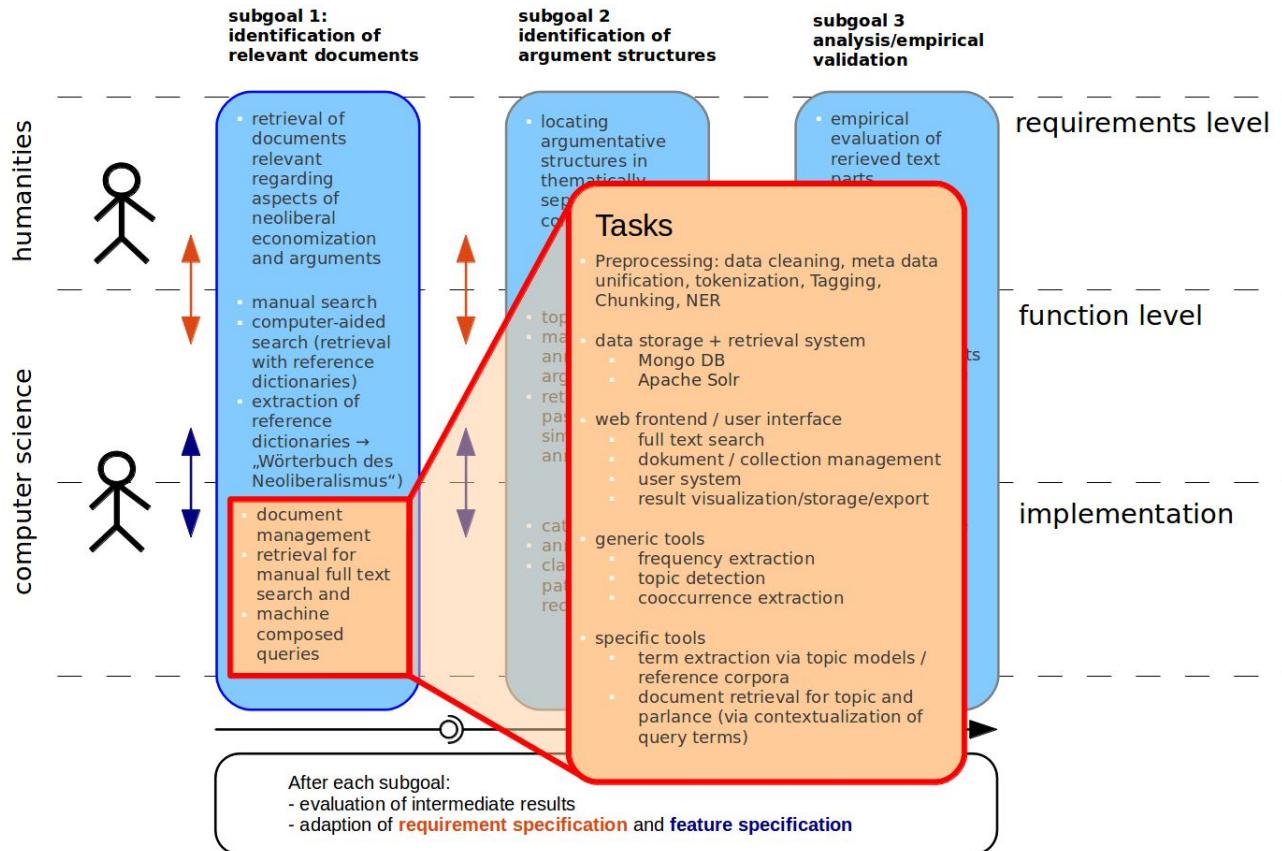


Abbildung 2 – Spezifikation konkreter Aufgaben für die Implementierung einzelner Teilaufgaben

Für diese drei Teilaufgaben werden Anforderungen von Seiten der Politikwissenschaftler formuliert und, dem „Lastenheft“ im Rahmen des Software Engineering vergleichbar, detailliert beschrieben. Auf der Implementierungsebene legen die beteiligten Informatiker fest, wie sie diese Anforderungen umzusetzen gedenken (vgl. „Pflichtenheft“). Geistes- und Sozialwissenschaftler als „informierte Anwender“ von Text Mining Verfahren und Informatiker als deren Entwickler sollten sich dabei in ihren Perspektiven auf die Umsetzung der Teilziele auf Ebene der funktionalen Spezifikation so nah wie möglich kommen. Das setzt voraus, dass die Politikwissenschaftler mögliche Analyse-Verfahren kennen und in ihren Grundlagen verstehen – zum Beispiel wenn entschieden werden muss, ob für ein induktives oder deduktives Forschungsdesign eher unüberwachte oder überwachte maschinelle Lernverfahren zum Einsatz kommen sollen. Die Informatiker wiederum benötigen ein Bewusstsein für die Anforderungen und Ziele der Politischen Theorie, um beispielsweise wie im Rahmen von ePol die Anforderung, dass es weniger um die Identifizierung klar abgegrenzter thematischer Bereiche für die Teilaufgaben 1 und 2 geht, als vielmehr um die zeitliche Veränderung eines Begründungsmodus, adäquat in ihrer Arbeit umsetzen zu können.

Für die Sicherung der Qualität eines solchen Forschungsprozesses liefert die Orientierung am Requirements Engineering ebenfalls einen nützlichen Rahmen. Die Aufteilung in wohldefinierte Teilziele ermöglicht jeweils die Evaluation von Zwischenergebnissen, bevor mit diesen die Analyse weiter fortgeführt wird. Auf Seiten der Informatik steht damit die Aufgabe, geeignete Evaluierungsverfahren für die jeweiligen Ergebnisse zu entwickeln, die wiederum mit Hilfe der Politikwissenschaftler durchgeführt werden müssen. Für Teilaufgabe 1 bedeutet das zum Beispiel die Annotation eines „Goldstandards“ von relevanten/nicht-relevanten Dokumenten in Bezug auf die Forschungsfrage mit denen die Qualität des Dokument-Retrieval bestimmt und optimiert werden kann. Systematische Evaluationen und semi-überwachtes Active Learning ermöglichen, dass Vertrauen in die computergestützten Analyseprozesse großer Datenmengen generiert und eine hohe Qualität in den oft quantifizierten Endresultaten sichergestellt wird.

Mit diesen Eigenschaften bilden die Methoden der Digital Humanities ein eigenartiges hybrides Gebilde zwischen dem positivistischen und dem interpretativen Paradigma empirischer Sozialforschung (Goldkuhl 2012). Beispielsweise dominiert bei der zunächst manuellen Identifizierung von neoliberalen Argumentationen im Teilschritt 2 des ePol-Projekts eindeutig eine hermeneutische Sicht auf die Daten. Aus dieser werden aber wiederum Sprachregelmäßigkeiten extrahiert, maschinell „gelernt“ und in einen quantifizierenden, eher positivistischen Analyseprozess überführt. Wie sehr ein solches Vorgehensmodell systematisiert und Forschungsprojekt-übergreifend festgelegt werden kann, ist noch längst nicht abschließend geklärt. Eine Debatte darüber, welche Anforderungen an die Qualitätssicherung des Forschungsprozesses, gerade im Hinblick auf die ungewöhnliche Kombination dieser oft widerstreitenden Wissenschaftsparadigmen zu stellen sind, muss noch viel intensiver geführt werden.

Wir sind jedoch der Auffassung, dass sich viele Erfahrungen aus dem ePol-Projekt zu einem Vorgehensmodell für größere Digital Humanities Projekte insgesamt generalisieren lassen. Die systematische Identifikation von (aufeinander aufbauenden) Teilzielen, ihre disziplinübergreifende Spezifikation auf verschiedenen Ebenen und ihre Untersetzung mit konkreten Aufgaben ermöglicht es, klare Erwartungen in die computergestützten Analysen zu entwickeln, den dafür notwendigen Arbeitsaufwand abzuschätzen und Kriterien für die Qualität der erreichten Ergebnisse bei den (Teil-)Zielen des Pro-

jets zu entwickeln. Insofern viele spannende geistes- und sozialwissenschaftlichen Fragen, die mit Hilfe großer Textkollektionen beantwortet werden könnten wohl kaum mit generischen Software-Lösungen bearbeitet werden können, sind weitere methodologische Innovationen zur Ermöglichung von Transdisziplinarität dringend notwendig.

Literatur

- Bohnsack, Ralf (2010): Rekonstruktive Sozialforschung. Einführung in qualitative Methoden. Ralf Bohnsack. 8. Aufl. Opladen, Farmington Hills, Mich: Budrich.
- Goldkuhl, Göran (2012): Pragmatism vs interpretivism in qualitative information systems research. In: European Journal of Information Systems 21 (2), S. 135-146.
- Kastens, Uwe / Büning, Hans Kleine (2008): Modellierung - Grundlagen und formale Methoden, Hanser.
- Keller, Reiner (2007): Diskursforschung. Eine Einführung für SozialwissenschaftlerInnen. 3., aktualisierte Aufl. Wiesbaden: VS Verlag für Sozialwissenschaften (Qualitative Sozialforschung, 14).
- Lemke, Matthias (2012): Die Ökonomisierung des Politischen. Entdifferenzierungen in kollektiven Entscheidungsprozessen. Discussion Paper Nr. 2. Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus. Hamburg; Leipzig.
- Schlieder/Wullinger (2010): Semantic Ageing of Complex Documents: A Case Study from Built Heritage Preservation, In: Informatik 2010, Service Science - Neue Perspektiven für die Informatik, Band 2, Gesellschaft für Informatik, 580-586.
- Wiedemann, Gregor; Lemke, Matthias; Niekler, Andreas (2013): Postdemokratie und Neoliberalismus – Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949-2011. Ein Werkstattbericht. In: Zeitschrift für politische Theorie 4 (1).

Vortrag 1

Ein Ei gleicht dem anderen

Automatische Analyse von Ähnlichkeit für historische Bildwissenschaften

Dr. Peter Bell, Computer Vision Group, Heidelberg Collaboratory for Image Processing (HCI), Universität Heidelberg, (AKDK)

Die umfangreichen digitalen Bildrepositorien von Museen, Bibliotheken, (Bild)archiven und Forschungsinstitutionen sind ein unschätzbarer Bestandteil der Infrastruktur für alle historischen Bildwissenschaften geworden. Dennoch hat der digitale Text im Netz einen deutlichen Vorsprung, da seine Zeichenfolgen leicht auffindbar sind, während der Inhalt des Bildes bislang nicht in gleicher Granularität durchsuchbar war. Die Speicherung des Bildes greift durch umfangreiche Verschlagwortung somit auf den Text zurück. Wenige Datenbanken beinhalten jedoch kaum mehr Informationen als Autorschaft, Datierung, Titel, Ort und Ikonographie. Komposition, Form, Lage der Objekte und Rezeptionszusammenhänge bleiben hingegen oft unberücksichtigt.

Die Heidelberger Kooperation von Kunstgeschichte und Computer Vision geht dieses Problem durch visuelle Suchverfahren, Objekterkennung und Szenenvergleiche an. Die bislang als Grundlagenforschung konzipierten Prototypen für Realien und Gesten in mittelalterlicher Buchmalerei und Architekturelemente sowie zu chinesischen Comics und assyrischer Keilschrift ergeben die Basis für Suchverfahren und Analyseinstrumente, die in allen historischen Bildwissenschaften einsetzbar sind und die teilweise bereits in einer Webapplikation getestet werden können.

Ansatzpunkt der interdisziplinären Zusammenarbeit ist die Frage nach Ähnlichkeit. Wie lassen sich verschieden enge Rezeptionsverhältnisse auffinden, vergleichen und abbilden? Diese Frage stellt sich in und über alle Kunstgattungen hinweg und entsprechend breit sollen die Untersuchungen hier für Malerei, Druckgrafik und Architektur vorgeführt werden. Ein facettenreiches Fallbeispiel stellen die vier illuminierten Ausgaben des Sachsenpiegels dar. Unterschiedliche Zeichenstile, verschiedener Produktions- und Erhaltungszustand und schließlich ein räumlich und zeitlich anderer Kontext führen zu Unterschieden zwischen den vier Illustrationsfolgen des 14. Jahrhunderts.

Die vorzustellenden Algorithmen können einerseits Szenen und Objekte in den verschiedenen Handschriften auffinden, indem diese in einem Codex markiert oder vom Nutzer hochgeladen werden. Die Treffermenge wird nach ihrer Ähnlichkeit angeordnet. Auf diese Weise lassen sich häufig wiedergegebene Objekte wie das Schwurreliquar, Kronen oder Wappen auffinden. Für die Anwendung auf Druckgrafik oder Architektur bedeutet dies, dass nicht nur identische Bilder, sondern auch Drucke von der gleichen Platte oder gleich gestaltete Gebäudeteile detektiert werden können. Davon zu unterscheiden sind Variationen, die gewisse Charakteristika des Originals teilen. So werden im Sachsenpiegel verschiedene liegende Figuren gefunden, deren Alter oder Geschlecht nicht mit der vom Nutzer markierten Person eines Greisen übereinstimmen. Dem Algorithmus gelingt es somit eine spezifische Körperhaltung zu erkennen, besonders erfolgreich gelingt dies für die Handgesten. In der Architektur können so variierende Bauteile, wie unterschiedlich gestaltete Kapitelle einer Ordnung oder von ihrer Form abweichende Baluster verschiedener Balustraden nach Ähnlichkeit geordnet werden. Heterogene Bilddatensätze können auf diese

Weise nach Rezeption durchsucht werden; etwa nach dem Bildzitat einer antiken Statur in der frühneuzeitlichen Malerei.

Im Gegensatz zu bereits verwendeten bag of visual words Ansätzen (Oxford, München) verwendet der vorgestellte Algorithmus auf HoG-Feature basierende Classifier mit prototypischen Negativbeispielen und während der Suche hinzugenommenen positiven Ergebnissen. Dies ermöglicht nicht nur das Auffinden von identischen oder sehr ähnlichen Partien, sondern auch von größeren Abweichungen.

Sind zwei ähnliche Szenen oder Objekte gefunden worden, lässt sich ein zweiter Analyseschritt anschließen, indem die Abweichungen genauer untersucht werden. Um die Konturen der Bilder zu vergleichen, werden die Transformationen errechnet, die nötig wären, um sie annähernd identisch werden zu lassen. Genauere Ergebnisse als herkömmliche Matching-Verfahren ergeben sich durch die Einteilungen von Gruppen gleicher Transformation. Diese vom Computer selbstständig definierten Kontursegmente geben nicht nur einen abstrakten Wert für Abweichung, sondern liefern auch Hinweise zum künstlerischen Prozess der Rezeption. So zeigen Zeichnungen nach Michelangelos Fresken in der Sixtinischen Kapelle, dass der Kopist recht detailliert einzelne Körperkonturen erfassen kann, jedoch Schwierigkeiten hat diese innerhalb der Proportionen des ganzen Körpers richtig zu lokalisieren. Im Fall des Sachsen Spiegels lassen sich so Szenen zwischen der Dresdner und Wolfenbüttler Version vergleichen oder auch die Varianz eines Objekts innerhalb einer Handschrift erkennen. Durch diese Vergleiche zeigt sich auch die funktionelle Motorik der Bildfiguren, deren Hauptartikulation Handgesten und Armhandlung sind. Die Ähnlichkeitsanalyse kann wichtige Informationen über die Genese der Bildwerke geben, sowohl im Hinblick auf technische Umsetzung wie auch den dahinter stehenden Auffassungen.

Erste automatische Vergleiche von frühneuzeitlichen Fassaden zeigen wie mit Hilfe eines Kantendetektor und dem Messen und Gewichten der auf einer Achse liegenden Konturen Strukturen und Rhythmen der Fassadengestaltung abgeleitet werden können. Eine klar strukturierte klassizistische Fassade lässt sich somit leicht von einem Rokoko-Lustschloss abgrenzen.

Die vorgestellten Algorithmen haben somit nicht nur den Zweck Bilder wieder aufzufinden, sondern rekonstruieren darüber hinaus Rezeptionsverhältnisse und analysieren die Ähnlichkeit mit einer für Menschen kaum durchführbaren Präzision. Diese Auswertung der visuellen Form steht komplementär zur um den semantischen Inhalt kreisenden Textannotation. Methodisch ergeben sich daraus Anknüpfungen an Stilkunde und Formanalyse zur Unterstützung von subjektiver Kennerchaft. Die enge Zusammenarbeit mit den Betreibern (kunst)historischer Bilddatenbanken und weiteren Projekten in den Digital Humanities ermöglicht die beschriebenen Such- und Analyseverfahren in die Infrastruktur der jeweiligen Fächer einzubinden.

Vortrag 2

Perspektiven der Forschung - PDF?

Digitale Bildwissenschaft zwischen gestern und morgen

Dr. Martin Raspe

Bibliotheca Hertziana - Max-Planck-Institut für Kunstgeschichte, Rom, (AKDK)

Mit etwas Verspätung, verglichen mit den textbezogenen Wissenschaften, sind die Bildwissenschaften im Verein der "digital humanities" angekommen. Trotz mancher Widerstände - nicht nur von Seiten der älteren Generation - ist der Computer aus der Kunstgeschichte, der Archäologie und ihren Nachbarwissenschaften nicht mehr wegzudenken. In allen Bereichen wird inzwischen computergestützt gearbeitet - von der Materialsammlung über die Auswertung bis hin zur Publikation und Dissemination. Was fehlt also?

Die Bildwissenschaft verwendet zwar Technologien von heute, arbeitet aber vielfach mit sehr traditionellen Methoden. Das betrifft alle Bereiche wissenschaftlichen Arbeitens. Die erheblichen epistemologischen Möglichkeiten, die das digitale Zeitalter bereitstellt, werden oft nicht einmal ansatzweise ausgeschöpft bzw. überhaupt als Positivum eingeschätzt.

Sinnbild dafür ist das PDF. Die Kunsthistorik nutzt heutzutage hochauflösende Scans und Fotografien, konsultiert Forschungsinformationen und "open linked data" im Internet und wendet digitale Analyseverfahren an. Wenn es aber an das Veröffentlichen der Ergebnisse geht, produziert man ein Textdokument im PDF-Format, das formal und inhaltlich nichts anderes ist als die digitale Repräsentation eines Buches oder Aufsatzes. Eigentlich hängt man immer noch an der guten alten Printpublikation. Die digitale Form wählt man nur notgedrungen, in erster Linie aus Kosten- oder Zeitgründen.

Gewiss hat ein PDF auch Vorteile gegenüber dem gedruckten Buch. Man kann es leichter überallhin mitnehmen, es ist durchsuchbar und kann sogar von Google analysiert und indexiert werden. Wenn es gut gemacht ist, kann man Textteile herauskopieren, vom Inhaltsverzeichnis aus zu den Kapiteln springen und sogar Internetlinks anklicken, doch darin erschöpft sich der Mehrwert. Das zusammengetragene und erarbeitete Wissen, die wissenschaftlichen Aussagen sind als solche nicht digital abrufbar, sondern bleiben letztendlich wie im traditionellen Buch "vergraben".

Abbildungen sind statisch. Sie werden in das Dokument hineinkopiert und mit jeder Kopie vervielfältigt. Wissenschaftliche Belege sind nur "von Hand" zu überprüfen und müssen auf herkömmliche Weise aufgesucht werden. Viele Vorteile der digitalen Informations- und Kommunikationswege werden verschenkt.

Ähnlich verhält es sich in anderen Bereichen der digitalen Bildwissenschaft. Überall werden traditionelle Verfahren ins Digitale übersetzt, ohne die neuen Möglichkeiten zu nutzen. Was ersetzt den Zettelkasten? Die private Datensammlung auf dem Laptop, die meist mit anderen Systemen inkompatibel ist und mit der Zeit veraltet oder gar unlesbar wird. Was ist aus dem wissenschaftlichen Nachschlagewerk geworden? Ein Datenbanksystem, das zwar online konsultierbar ist, aber nur nach denjenigen Metadaten abgefragt werden kann, die im System vorgesehen sind. Inhaltlich und funktional bleibt es genauso abgeschlossen wie die Printausgabe.

Der Vortrag beleuchtet anhand verschiedener Beispiele und Anwendungen den derzeitigen Stand der digitalen Technologie und die methodologischen Konsequenzen für ein bildorientiertes Fach wie die Kunstgeschichte.

Vortrag 3

Das Museum als digitaler Lernort

Georg Hohmann M.A.

Deutsches Museum München, (AKDK)

Ein zentraler Praxisort der Kunsthistorik ist das Museum, das definitionsgemäß eine ganze Reihe von Aufgaben zu erfüllen hat, deren gegenseitige Abwägung nicht immer leicht ist. Zweifellos wächst zurzeit die Aufmerksamkeit, die dem Museum als Lernort gewidmet wird, und in diesem Zusammenhang steigen die Erwartungen an digitale Verfahren und Techniken.

Wie kann in diesem Szenario die Haltung der Museen selbst beschrieben werden, und gibt es strukturelle Unterschiede in der Selbstwahrnehmung der einzelnen Institutionen? Hier lassen sich extrem unterschiedliche Haltungen beobachten. Auf der einen Seite finden sich Positionen, die auf die Auratik der Begegnung des Publikums mit den ausgestellten Artefakten bauen, während auf der anderen Seite vielfältige digitale Medien eingesetzt werden, um das durch die Sammlungsgegenstände repräsentierte Thema in weiteren Facette zu vermitteln. Stehen diese unterschiedlichen Positionen mit dem Kunstcharakter der Artefakte in Zusammenhang? Lassen sich wissenschaftsgeschichtliche Traditionen rekonstruieren und benennen, die die aktuelle Haltung zu digitalen Vermittlungstechniken präfigurieren?

In dem Vortrag sollen - aus der langjährigen Praxiserfahrung in den digitalen Abteilungen verschiedener Museumstypen heraus - unterschiedliche Strategien bei der Etablierung des Museums als digital aufgestelltem Lernort vorgestellt werden.

Gez. Stephan Hoppe

Für schnelle Kontaktaufnahme: 0172 – 36 37 836



LMU □ Geschwister-Scholl-Platz 1 □ 80539 München

An das Programmkomitee

Prof. Dr. Stephan Hoppe

Telefon +49 (0)89 2180- 3500
Telefax +49 (0)89 2180- 5316

email@stephan-hoppe.de

<http://stephan-hoppe.de>

<http://www.kunstgeschichte.uni-muenchen.de/ifk/index.html>

Postanschrift
Geschwister-Scholl-Platz 1
80539 München

Ihr Zeichen, Ihre Nachricht vom

Unser Zeichen

Dresden, 31. Dezember 2013

Vorschlag für eine Sektion für die Tagung „Digital Humanities im deutschsprachigen Raum (DHd)“

Bedrohte Besitzstände, verlorene Werte? Die Geisteswissenschaft von der Kunst und die neuen digitalen Verfahren.

Sektionsleitung: Prof. Dr. Stephan Hoppe

LMU München, Institut für Kunstgeschichte, in Verbindung mit dem Arbeitskreis Digitale Kunstgeschichte (AKDK)

Es ist kein Geheimnis, dass gerade in den Geisteswissenschaften die neuen digitalen Technologien und Verfahren nicht nur als willkommene Erweiterungen der Möglichkeiten angesehen werden. So hat z.B. der Germanist Philipp Theisohn jüngst die Verfolgung digitaler Suchstrategien in die Nähe der Täuschung und des Plagiats gerückt: „... Im Zuge der Digitalisierung ist es ein Leichtes geworden, dieses unsichtbare Kapital zu simulieren und mit geliehener Gelehrsamkeit zu handeln, indem man jene langen Wege vergeblichen Suchens und Lesens den Computer gehen lässt und sich dann nur noch mit den „Treffern“ befasst. ... Die Probleme, die diese Konstellation in Forschung und Lehre verursacht (und der Plagiarismus ist hierbei eher eines der kleineren Probleme), sind unübersehbar und geben einen ersten Hinweis darauf, was „digitale Verfügbarkeit“ in der Wissenschaft letztlich auch bedeuten kann: Selbsttäuschung und Blenderei ...“ (Philipp Theisohn 2012)

Hier handelt es sich um eine Bewertung und Position, die eigentlich alle Geisteswissenschaftler angehen müsste, da inzwischen kaum jemand vollständig auf die Nutzung einer Suchmaschine verzichten darf und somit zumindest in die Nähe der angesprochenen Gefahren gelangen kann. Hier geht es grundsätzlich nicht um die Beurteilung der Tauglichkeit konkreter technischer Verfahren im Hinblick auf ein bestimmtes wissenschaftliches Erkenntnisinteresse, sondern um die Formulierung und Einforderung bestimmter Haltungen und Vorgehensweisen als für gute geisteswissenschaftliche Praxis konstituierend. Ethische Maximen sind sinnvoll und notwendig; ebenso notwendig ist aber auch ihre Reflexion und rationale und transparente Fundierung jenseits von Polemik und habituellem Zwang.

Es ist jedoch zu befürchten, dass auch in der Praxis einer der klassischen geisteswissenschaftlichen Disziplinen wie der Kunstgeschichte mit vielen digitalen Neuerungsangeboten in vielen Bereichen auf einer eher unausgesprochenen und latenten Ebene umgegangen wird. Nicht immer ist z.B. klar, ob die in verschiedenen Kreisen des Faches Kunstgeschichte tatsächlich zu beobachtende Zurückhaltung gegenüber dem aktuellen technischen Stand auf nur partieller Kenntnis, technologischen Bewertungen, ökonomischen Abwägungen oder moralischen Überzeugungen beruht. Zudem fehlt oft die Erarbeitung bzw. Adressierung plausibler und hinreichend komplexer Zukunftsszenarien, die zumindest hypothetisch Auskunft über die Stellung kunsthistorischer Phänomene in der mittelfristigen Zukunft geben könnten.

Welches wären überhaupt die problematischsten Herausforderungen, die sich durch den vermehrten Einsatz digitaler Verfahren im Bereich der Kunstgeschichte ergeben würden? Ist es das gesamtgesellschaftlich zweifellos zu den Kernherausforderungen gehörende Feld des Datenschutzes und der Privatsphäre? Das universelle Urheberrecht und die Abwägung mit anderen Rechtsgütern? Oder handelt es sich um fachspezifischere Herausforderungen wie den gefürchteten oder erhofften Ersatz typischer geisteswissenschaftlicher Arbeitsverfahren durch Maschinen, die Auflösung eines als ethisch wertvoll erachteten ästhetischen Kanons oder den schlechenden Verlust von einstigen Monopolen der Welterklärung, Kennerschaft und Ordnung ihrer Artefakte?

In der vorgeschlagenen Sektion sollen – und dies ist weder abschließend oder allumfassend gemeint – drei Bereiche signifikanter Neuerungsansprüche im Bereich kunstwissenschaftliche Praxis diskutiert werden. In dem Vortrag von Peter Bell wird der aktuelle Stand der maschinellen Bilderkennung vorgestellt und die Auswirkungen auf traditionelle kunsthistorische Arbeitsszenarien angedeutet. Auch der Vortrag von Martin Raspe stellt die Frage nach nun möglichen neuen Verfahren im Fach, hier konkret im Bereich des Fachdiskurses und allgemein der fachbezogenen Kommunikation nach dem Ende der Gutenberggalaxie. In dem Dritten Beitrag schließlich soll ebenfalls eine zentrale kommunikative Schnittstelle des Faches beleuchtet werden, und zwar das Museum mit seinen nun schon vielfältigen Erfahrungen auf dem Gebiet digitaler Vermittlungsangebote von Kunst.

Ziel der Sektion ist nicht die Generierung einer größeren Akzeptanz digitaler Verfahren oder gar eine endgültige Klärung der oben angesprochenen Bedenken, sondern die Beförderung eines expliziten und konstruktiven Diskurses in einem konkreten geisteswissenschaftlichen Fach und seinen Nachbardisziplinen.

LitSOM. Kartierung russischer Gegenwartsliteratur

Gernot Howanitz, Lehrstuhl für Slavische Literaturen und Kulturen, Universität Passau
Helmut A. Mayer, Fachbereich Computerwissenschaften, Universität Salzburg

Zusammenfassung

Das literarische SOM (LitSOM) wendet selbstorganisierende Karten (Self-Organizing Maps, SOM) und Learning Vector Quantization (LVQ) auf russische Literatur an. Das SOM wird dafür eingesetzt, um eine Karte zeitgenössischer russischer Romane zu erstellen, die es Literaturwissenschaftlerinnen und Literaturwissenschaftlern erlaubt, Beziehungen zwischen den Romanen zu untersuchen. LitSOM ist eine ‚Distant reading‘-Technik. Die Qualität der Karten wird sowohl subjektiv aus der Sicht der Literaturwissenschaft, als auch objektiv, d.h. in einem ‚klassischen‘ Problem des Textmining, nämlich der Klassifizierung von Autorinnen und Autoren, bestimmt. Um dies zu erreichen, wird der SOM-Algorithmus durch den LVQ-Algorithmus ergänzt.

1. Einleitung

1.1 Überblick

Ein Hauptziel dieses Beitrags ist die Implementierung eines Systems für computerunterstützte Textanalyse, das sogenannte *Literary SOM* (LitSOM). Darüber hinaus zeigen wir, wie Literaturwissenschaftlerinnen und Literaturwissenschaftler dieses System für ihre Forschungen einsetzen können. In unserer Beispielanwendung haben wir jeweils 15 Romane von acht verschiedenen Autorinnen und Autoren der russischen Gegenwartsliteratur kartiert, um die Nützlichkeit von quantitativen Methoden für die slavistische Literaturwissenschaft zu demonstrieren. Zwar gibt es eine lange russische Tradition quantitativer Zugänge zur Literatur, diese ist allerdings etwas in Vergessenheit geraten. So hat der bedeutende russische Mathematiker Andrej Markov 1913 ein Paper publiziert [1], in dem er das Konzept der Markovkette demonstriert. Markovketten erlauben es, Ereignisse zu modellieren, die nacheinander stattfinden. Heutzutage werden sie in verschiedensten Feldern angewandt, beispielsweise in den Wirtschaftswissenschaften oder in der Physik, und sie spielten auch eine Schlüsselrolle in Claude Shannons grundlegender Monographie zur Informationstheorie [2]. Trotz dieser vielfältigen Anwendungsmöglichkeiten hat Markov im Jahr 1913 den Versroman ‚Eugen Onegin‘ (1833) zu seinem Studienobjekt gemacht. Dieses Beispiel zeigt, wie stark die Verbindung zwischen Mathematik und Literatur im Russland des frühen 20. Jahrhunderts war.

1.2 Methodologie

LitSOM basiert auf sogenannten selbstorganisierenden Karten (Self-Organizing Maps, SOM) [3]. Ein SOM ist perfekt geeignet für unstrukturierte Daten und unvollständige Information, weil es hochdimensionale Probleme vereinfachen und für den Menschen leicht verständlich darstellen kann. Deshalb eignen sich SOM sehr gut für Data Mining [4]. Ein SOM kann dazu verwendet werden, eine große Anzahl von Texten zu clustern und sie auf einer zweidimensionalen Karte anzuzeigen. Das sogenannte WEBSOM [5] clustert beispielsweise Newsgroup-Postings nach ihrem Inhalt. Das LitSOM funktioniert ähnlich, arbeitet aber mit Romanen anstelle kurzer Nachrichten. Es erstellt eine Karte, die die Abstände zwischen einzelnen Romanen darstellt – je näher, desto ähnlicher. Dieser Text-Mining-Ansatz liefert Literaturwissenschaftlerinnen und Literaturwissenschaftlern eine automatische Visualisierung von Beziehungen zwischen unterschiedlichen Texten. Nach Franco Moretti ist LitSOM ein Werkzeug für ‚distant reading‘ [6], also für eine Mischung aus der klassischen literarischen Textanalyse („close reading“) und dem Querlesen von Texten [7].

2. Vorarbeiten

2.1 Implementierung des SOM

Alle Bestandteile von LitSOM (SOM, Feature Extraction basierend auf Wortfrequenzen und eine Visualisierung mittels der Unified Distance Matrix [8]) wurden in Java implementiert. Für die Feature Extraction haben wir Sergej Sharovs Liste der 5000 häufigsten russischen Wörter verwendet [9]. Der Feature-Vektor wurde dann wie folgt zusammengestellt: Für jeden Roman wurden die Wörter aus der Sharov-Liste gezählt und jeweils durch die Gesamtanzahl der Wörter in diesem Roman dividiert. Dies gewährleistet, dass die einzelnen Feature-Vektoren untereinander vergleichbar bleiben.

2.2 Setup der Experimente

Verschiedene Längen des Feature-Vektors wurden getestet: 5, 10, 25, 30, 40, 50, 75, 100, 125, 150, 175 und 200 Features. Um den Einfluss verschiedener Wortarten auf die resultierenden Karten zu untersuchen, wurde Sharovs ursprüngliche Liste modifiziert; Versionen rein mit Nomen und Verben sowie eine Kontrollgruppe mit allen anderen Wörtern wurde erstellt. Aufgrund eigener Testreihen haben wir uns für ein SOM aus 108 Neuronen in einem hexagonalen 9×12 Gitter entschieden. Die Lernrate $\alpha(0)$ wurde auf 0.5 gesetzt und dann wie folgt verringert: $\alpha(t + 1) = \alpha(t)/(1 + \alpha(t))$. Der anfängliche Nachbaschaftradius wurde mit 2.5 festgelegt. Nach jedem Zyklus wurde dieser Radius um 0.0005 verringert. Nach einer unüberwachten SOM-Phase mit 3000 Zyklen folgte eine überwachte LVQ-Phase mit 1000 Zyklen. Insgesamt wurden 4800 Experimente durchgeführt und ebensoviele Karten erstellt.

Tabelle 1

Rang	Feature-Vektor	Korrekt identifiziert	LVQ-Genauigkeit	1NN-Genauigkeit
1	150 Verben	103	85,83%	92,50%
2	100 Nomen	102	85,00%	95,83%
3	100 Verben	101	84,16%	91,67%
4	200 Sharov	100	83,33%	90,00%
5	175 Nomen	99	82,50%	95,00%
6	100 Sharov	98	81,67%	90,00%
6	150 Sharov	98	81,67%	90,83%
6	125 Nomen	98	81,67%	93,34%
6	125 Verben	98	81,67%	91,67%
10	75 Nomen	97	80,83%	94,16%
10	40 Nomen	97	80,83%	89,16%
10	175 Verben	97	80,83%	92,50%

3. Resultate

3.1 Klassifizierung von Autorinnen und Autoren

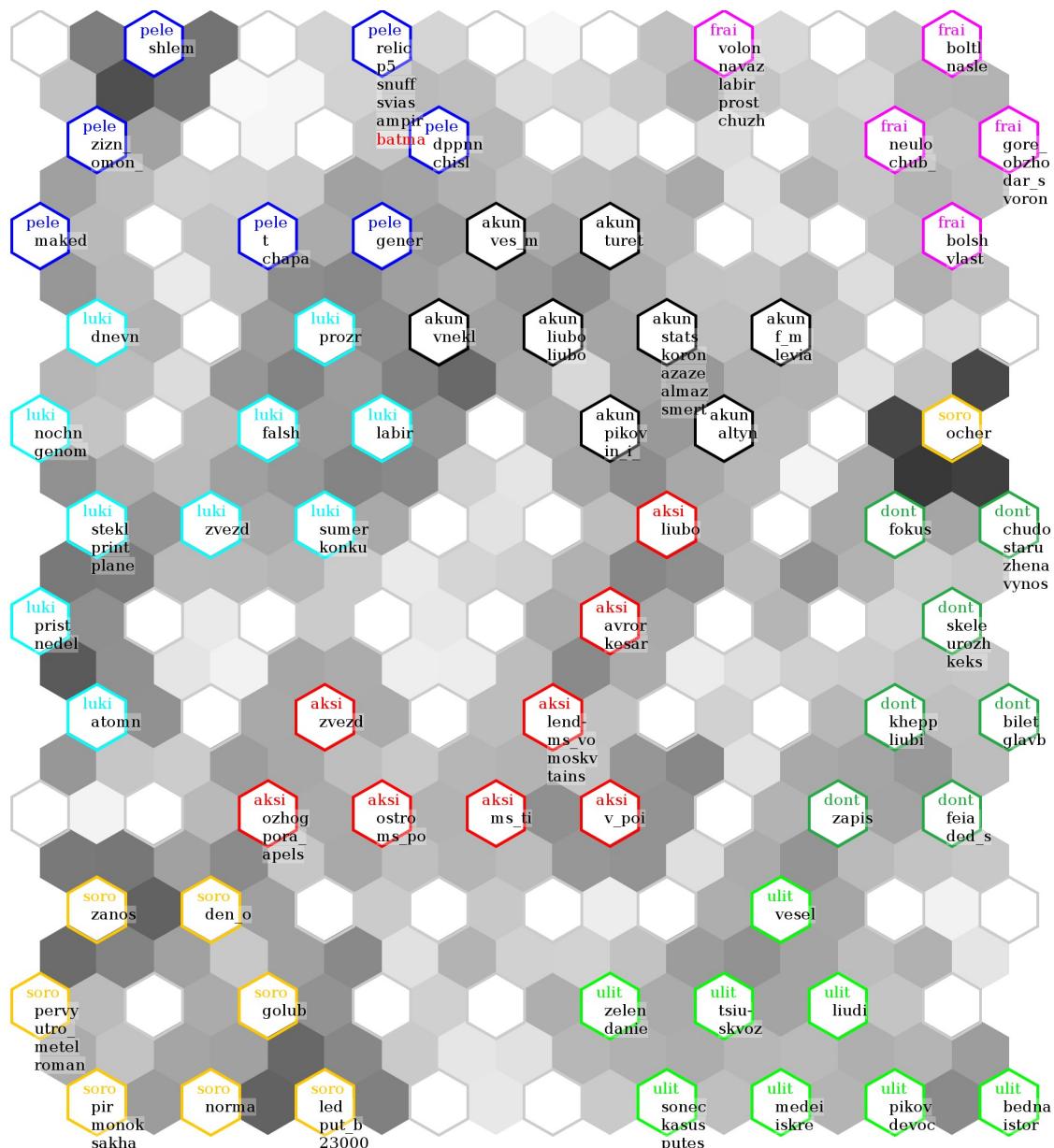
40 verschiedene Konfigurationen und Leave One Out Cross Validation (LOOCV) resultierten in einer Gesamtanzahl von 4800 verschiedenen Experimenten. Die besten Resultate dieser 4800 SOM/LVQ-Läufe sind in Tabelle 1 angeführt. Das beste Resultat – 86% richtig erkannt – wurde mit

einer Liste von den 150 häufigsten Verben als Feature-Vektor erzielt. Diese Resultate legen den Schluss nahe, dass die von LitSOM produzierten Karten die Verteilung der 120 Romane tatsächlich widerspiegelt. Mit einem 1NN-Klassifizierer, der als Kontrolle fungierte, wurde sogar eine Genauigkeit von 96% erreicht.

3.2 *U-Matrix*

Die Qualität der Karten kann nur subjektiv bestimmt werden. Deshalb präsentieren wir hier eine Karte samt Interpretation als Beispiel. Im Allgemeinen ist anzumerken, dass zwischen einzelnen Karten durchaus Unterschiede festzustellen waren, allerdings glichen sich die Karten trotzdem meist in ihrer grundlegenden Struktur.

Grafik 1: U-Matrix-Visualisierung des SOM für Viktor Pelevins „Ananaswasser für eine feine Dame“



Grafik 1 zeigt die U-Matrix, die das LitSOM für Viktor Pelevins Roman „Ananaswasser für eine feine Dame“ nach den SOM/LVQ-Läufen darstellt. Diese Karte wurde basierend auf einem Pattern-Vektor mit den 100 häufigsten Nomen erstellt. Pelevins Roman diente als unbekannter Test-Text,

d.h. nach dem Training mit den 119 anderen Romanen wurde dieser Text – korrekt – klassifiziert. Wie man sieht, weist LitSOM sehr gut auf Romane hin, die eher untypisch für die jeweilige Autorin oder den jeweiligen Autor sind. Beispiele dafür sind Vladimir Sorokins „Die Schlange“ (gekennzeichnet durch „ocher“). Auch die jeweiligen Relationen unterschiedlicher Autorinnen und Autoren zueinander sind nachvollziehbar, so liegen die beiden Fantasy-Autoren Sergej Lukjanenko und Viktor Pelevin nebeneinander. Unser letzter Test fand sozusagen unter realistischen Bedingungen statt: Pelevins neuester Roman „Batman Apollo“ wurde am 8. März 2013 publiziert, nachdem der Großteil unserer Experimente bereits abgeschlossen war, damit hat er auch nicht Eingang in das ursprüngliche Textkorpus gefunden. In Grafik 1 findet man „Batman Apollo“ („betman“ in rot) gleich neben weiteren Pelevin-Romanen jüngeren Datums, vor allem auch „Empire V“ („ampir“). Blättert man diese Romane durch, erfährt man, dass „Batman Apollo“ die Fortsetzung von „Empire V“ ist.

4. Diskussion

Die Ergebnisse der Klassifizierungsexperimente mit LVQ und 1NN sind sehr gut, vor allem in Anbetracht der Tatsache, dass literarische Texte sehr komplex sein können. Wortfrequenzen erlauben es, zwischen Romanen unterschiedlicher Autorinnen und Autoren zu differenzieren. Mit der Liste der 150 häufigsten Verben konnten 103 von 120 Romanen (86%) korrekt ihren jeweiligen Autorinnen und Autoren zugeordnet werden. Damit wurde empirisch belegt, dass LitSOM die Beziehungen zwischen Texten unterschiedlicher Autorinnen und Autoren sinnvoll darstellen kann. Die Visualisierung mittels U-Matrix, die LitSOM auch zur Verfügung stellt, erlaubt es, die Relationen zwischen unterschiedlichen Texten in einfacher Form darzustellen. Unsere subjektive Bewertung der Karten zeigte, dass die Wahl der Features großen Einfluss auf die visuelle Qualität der Karten hat. So waren die Cluster einzelner Autorinnen und Autoren bei auf der Nomen-Liste basierenden Karten am besten voneinander abgetrennt. Die Verben-Liste wiederum war für das Klassifizierungsexperiment besser geeignet, optisch waren die Karten allerdings weniger klar strukturiert. Im Allgemeinen eignen sich die Karten vor allem dazu, Texte zu finden, die für einen Autor oder eine Autorin untypisch sind bzw. die dem Stil einer anderen Autorin oder eines anderen Autors ähneln. Auch innerhalb eines Clusters lassen sich interessante Schlüsse hinsichtlich der Texte ziehen, so gibt es häufig Unterschiede zwischen noch in der Sowjet-Ära geschriebenen Texten und späteren, post-sowjetischen.

LitSOM kann in vielerlei Hinsicht verbessert werden; bei den Feature-Vektoren sind noch viele weitere Kombinationen denkbar, die durch Feature-Selection-Algorithmen bestimmt werden könnten. Weiters ist es denkbar, die visuellen Karten automatisiert durch Bildverarbeitungs-Algorithmen zu vergleichen, um eine objektivere Beurteilung zu erreichen. Auch der Einfluss der SOM-Parameter, etwa unterschiedlicher Kartengrößen, auf die visuelle Qualität der Karten ist noch nicht hinreichend untersucht. Weitere wertvolle Einblicke könnten durch Einbeziehung weiterer Texte aus unterschiedlichen literarischen Epochen gewonnen werden. Gleichzeitig würden all diese Experimente helfen, mehr Erfahrung im Umgang mit den Karten zu gewinnen. Diese Erfahrung ist sehr wichtig, denn schlussendlich kann nur ein Mensch die Karten interpretieren und als Ausgangspunkt für weitere Überlegungen nutzen – ein Prozess, der ‚distant reading‘ und ‚close reading‘ verbindet.

Quellen

1. Markov, A. 1913. Primer statisticheskogo issledovaniia nad tekstom ‚Evgeniia Onegina‘, illiustriruiushchii sviaz ispytanii v tsep’. *Izvestiia Imperatorskoi Akademii Nauk* 7.3.
2. Shannon, CE. and Weaver, W., 1949. *The Mathematical Theory of Communication*. University of Illinois Press, Illinois.
3. Kohonen, T. 2001. *Self-Organizing Maps*. Berlin.
4. Lagus, K. et al. 1999. WEBSOM for Textual Data Mining. *Artificial Intelligence Review* 13 (5/6).

- <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.5452> [accessed 17 March 2013].
5. Kohonen, T. et al. 2000. Self Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks* 11 (3), 574–585. <http://lib.tkk.fi/Diss/2000/isbn9512252600/article7.pdf> [accessed 14 March 2013].
 6. Moretti, F. 2000. Conjectures on World Literature. *New Left Review* 1, 54-68. <http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature> [accessed 30 November 2013].
 7. Kirschenbaum, M. 2007. The Remaking of Reading: Data Mining and the Digital Humanities. *National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation.* <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.959&rep=rep1&type=pdf> [accessed 9 May 2013]
 8. Kohonen, T. 2001. *Self-Organizing Maps*. Berlin, 165f.
 9. Sharov, S. 2001. Chastotnyi slovar'. *RosNII II Website*. <http://www.artint.ru/projects/frqlist.php> [accessed 12 March 2013].

Abstract für die Jahrestagung der Digital Humanities im deutschsprachigen Raum, 25.-28.3.2014,
Passau *Digital Humanities – methodischer Brückenschlag oder “feindliche Übernahme”?*

Ben Kaden (TU Berlin / ben.kaden@tu-berlin.de)

Schwerpunkt: Digitale Infrastrukturen für die Geisteswissenschaften

Ein Integriertes Monitoring in digitalen Forschungsräumen als Vermittlungsakteur in heterogenen Akteurskonstellationen. Das Beispiel TextGrid.

Die Entwicklung und Verbreitung komplexer digitaler Werkzeuge und Forschungsinfrastrukturen u.a. im Bereich der *Digital Humanities* erfordert spezifische Beurteilungs-, Entscheidungs- und Handlungskompetenzen, die für die beteiligten Akteure, insbesondere die FachwissenschaftlerInnen, wenigstens teilweise neu sind. Mehr als zuvor müssen sie sich mit digitalen Programm- und Verarbeitungsstrukturen, mit (Meta)Datenschemata und den digitalen Repräsentationsmöglichkeiten für ihre Forschung auseinandersetzen. Dies betrifft insbesondere die editionsphilologische Forschung, die als zentrale Zielgruppe der virtuellen Forschungsumgebung TextGrid gelten kann.

Hier bzw. generell in digitalen Infrastrukturen wirken drei Kompetenzbereiche zusammen:

a) die Fachwissenschaft, b) die Informatik und c) die Bibliothekswissenschaft.

Mit diesen Bereichen sind naturgemäß jeweils besondere Anforderungen an und Blickwinkel auf die Forschungsinfrastruktur verbunden, die sich zum Teil erheblich voneinander unterscheiden. Eine zentrale Herausforderung solch heterogener Akteurskonstellationen liegt daher in der Übersetzungsleistung zwischen diesen drei Bereichen bzw. den jeweiligen Akteuren.

So muss erstens ein erheblicher Aufwand betrieben werden, damit die Bedarfe der FachwissenschaftlerInnen und die informationstechnischen Möglichkeiten sinnvoll und erfolgreich aufeinander bezogen werden können. Oft schätzen WissenschaftlerInnen die technische Umsetzung ihrer spezifischen Anforderungen und besonders ihrer Workflows im Kontext eines Forschungsprojektes in einen digitalen Forschungsraum, wie ihn die Infrastrukturen anbieten möchten, nicht als befriedigend erfüllt ein. Die Folge ist eine zurückhaltende bis ausbleibende Annahme des Angebots.

Zweitens sehen sich EntwicklerInnen im Gegenzug mitunter durch die Anforderungen der FachwissenschaftlerInnen herausgefordert, wenn sie deren Ansprüche und Workflows in eine technische bzw. Programmstruktur zu überführen versuchen. Dies ist besonders dann der Fall, wenn die EntwicklerInnen keinen zusätzlichen fachwissenschaftlichen Hintergrund besitzen. So erschwert ein ungenaues Verständnis der formulierten Anforderungen an die Forschungsinfrastrukturen auf Seiten der EntwicklerInnen eine passgenaue Umsetzung in der Technik erheblich. Bei einer in der Regel begrenzten Ressourcenlage sind Fehlentwicklungen besonders problematisch. Zudem gilt es bei Forschungsinfrastrukturen die Nachhaltigkeit auch bei sich verändernden technischen Rahmenbedingungen zu berücksichtigen.

Den dritten Aspekt stellen die Anforderungen und Wünsche der archivierenden Institutionen (zum Beispiel Bibliotheken oder auch Rechenzentren) bezüglich der (a) Langzeitarchivierung und -verfügbarhaltung sowie der (b) Interoperabilität der publizierten Inhalte und Daten dar. Diese müssen mit dem technisch Machbaren (Informatik) und dem von den FachwissenschaftlerInnen Gewünschten integriert werden. Auch hier spielt ein möglicher Wandel der technischen

Rahmenbedingungen (Speicher- und Abbildungstechnologie, etc.) eine große Rolle. Weiterhin müssen die rechtlichen Bedingungen der Datenhaltung und –nutzung und die entsprechenden Wünsche der WissenschaftlerInnen koordiniert werden.

Dass eine entsprechende Abstimmung der zentralen Anforderungskomplexe , nämlich die Wissenschaftsworkflows und die Archivierung und Zugänglichmachung der Inhalte mit den Umsetzungsoptionen in der Technik nur über eine ständige Kommunikation zwischen den jeweiligen Akteuren funktionieren kann, ist angesichts der Komplexität der Aspekte offensichtlich. In der Praxis läuft diese erfahrungsgemäß nicht immer von selbst und nicht immer reibungslos ab.

Als Lösung könnte an dieser Stelle ein weiterer Akteur wirksam werden, der alle drei Perspektiven gleichermaßen im Blick behält, analysiert und zugleich die notwendigen Verständigungsprozesse aktiv koordiniert, moderiert und unterstützt. Entscheidend ist dabei, dass dieser Akteur sowohl von den anderen Akteuren, also den FachwissenschaftlerInnen, denen der technischen Entwicklung und dem technischem Betrieb sowie den Akteuren der Archivierung und Zugänglichmachung in dieser spezifischen Rolle akzeptiert und ernst genommen wird.

Im Projekt TextGrid wird im Rahmen der dritten Projektphase für diese Aufgabe ein so genanntes Integriertes Monitoring entwickelt. Dessen Aufgabenbereich lässt sich abstrakt beschreiben als

Beobachtung und Analyse der Wechselwirkungen zwischen den WissenschaftlerInnen, der Technik und den Inhalten in einer digitalen Forschungsinfrastruktur mit dem Ziel, Nutzung und Betrieb sowie weitere Entwicklungen des Angebots zu begleiten und bei Störungen und Problemen möglichst frühzeitig mit Steuerungsmaßnahmen einzutreten.

Das Angebot, der Betrieb und die Nutzung von Wissenschaftsdienstleistungen wie virtuellen Forschungsumgebungen werden dabei grundsätzlich als soziales Geschehen verstanden. Bei diesem interagieren heterogene Akteure mit unterschiedlichen Interessen und Kompetenzen vor dem Hintergrund eines verbindenden Ziels, nämlich dem reibungsarmen, zeitgemäßen und nachhaltigen Funktionieren digital gestützter Forschung.

Die Präsentation arbeitet die Konstellation sowie die dabei häufig auftretenen Verständigungsprobleme exemplarisch heraus und erläutert, wie das beim Projekt TextGrid derzeit in Entwicklung befindliche Konzept eines Integrierten Monitorings diesbezüglich ausgleichend und steuernd aktiv werden soll.

Da davon auszugehen ist, dass die im Bereich des gegenseitigen Verstehens liegenden Schwierigkeiten nahezu zwangsläufig bei fast allen derartigen Projekten auftreten und daher für die *Digital Humanities* als typisch zu bezeichnen sind, lässt sich das hier entwickelte Konzept des Integrierten Monitorings idealerweise als Prototyp verstehen, von dem ausgehend entsprechende Lösungsoptionen auch in anderen Kontexten als dem von TextGrid entwickelt werden können.

(Berlin, Dezember 2013)

eIdentity – Werkzeuge zur Erschließung und Exploration von Textdaten

Cathleen Kantner¹, Fritz Kliche², Jonas Kuhn³

¹Institut für Sozialwissenschaften
Universität Stuttgart

²Institut für Informationswissenschaft und Sprachtechnologie
Universität Hildesheim

³Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

Im Rahmen des BMBF-Verbundprojekts *eIdentity* arbeiten wir aktuell an einem mehrsprachigen Korpus von Zeitungstexten über Kriege und humanitäre militärische Interventionen aus unterschiedlichen Medienarchiven im Umfang von ca. 700.000 Dokumenten. Dieses Korpus wird aus einer politikwissenschaftlichen Perspektive daraufhin untersucht, welche kollektiven Identitäten, beispielsweise *europäische*, *nationale* oder *religiöse Identitäten*, im Zusammenhang mit internationalen Krisen ausgedrückt, beschworen oder kritisiert werden.

Im Projekt *eIdentity* entwickeln wir dazu 1) eine *Explorationswerkbank* für die Aufbereitung und das Management von potentiell heterogen strukturierten Textdaten für die Verwendung sprachtechnologischer Werkzeuge sowie 2) unseren *Complex Concept Builder*, ein System interaktiver Tools zur inhaltlichen Arbeit mit dem Korpus.

1) Explorationswerkbank

Wissenschaftler in den Digital Humanities stehen bei der Nutzung von zuvor nicht bearbeiteten elektronischen Textdaten vor dem Problem, dass es bislang keine benutzerfreundliche Software für die komplexen Aufgaben der Korpuserstellung und -aufbereitung, der Strukturierung von Texten und zugehörigen Metadaten sowie sich anschließende Aufgaben der Samplebereinigung und des Datenmanagements gibt. Doch erst wenn all diese Aufgaben erledigt sind, kann die eigentliche Textanalyse beginnen. Mit der *Explorationswerkbank* entwickeln wir kombinierbare Werkzeuge, die Wissenschaftlern in den Digital Humanities dabei helfen, diese massive Hürde zu Beginn ihrer eigentlichen empirischen Forschung zu bewältigen.

Allerdings stellen verschiedene Wissenschaftler in den Digital Humanities ganz unterschied-

liche Anforderungen an das Datenmanagement. Aus der Perspektive der jeweiligen fachlich-spezifischen Forschungsfrage sind eine Vielzahl von Entscheidungen zu treffen. Unser Ansatz besteht folglich darin, wo möglich den Anwendern selbst die Steuerung der Verarbeitungsschritte zu ermöglichen, so dass sie dazu nicht auf die Mitwirkung der Werkzeugentwickler angewiesen sind. Die Nutzer können zu verschiedenen Verarbeitungsschritten eigene Zielwerte festlegen. Dieses Konzept umfasst die Import-Werkzeuge, die Werkzeuge zur Datenbereinigung und -Filterung, die Auswahl verwendeter sprachtechnologischer Werkzeuge zum Auffinden relevanter Texte und Textstellen sowie Werkzeuge anhand des gewünschten Typs von Output.

Bereits bei der Korpuserstellung finden wir auf der einen Seite digitale Daten in unterschiedlichen Formaten und Datenstrukturen vor. Auf der anderen Seite lassen sich sprachtechnologische Werkzeuge meist nicht unmittelbar auf beliebiges Datenmaterial anwenden. Dies gilt besonders für die Analyse großer Datenmengen. Die *Explorationswerkbank* bringt beide Aufgaben zusammen. DH-Wissenschaftler können damit ihr rohes Datenmaterial aus verschiedenen Quellen aufbereiten und Texte und Metadaten in einem Repository ablegen, das anschließend die Einbindung sprachtechnologischer Werkzeuge erlaubt.

Für den Import roher Daten stellen wir eine Wizard-Funktion bereit. Die Anwender importieren zunächst ein einzelnes Dokument in ein „Vorschau-Fenster“ und erstellen anschließend Regeln, um im Dokument Metadaten und Textstrukturbasteine zu definieren. Daraufhin wenden die Import-Werkzeuge die Regeln auf die Dokumente der Datenquelle an (Generalisierung). Zur weiteren Verarbeitung entwickeln wir Werkzeuge für die Bereinigung des erstellten Samples. Die Bereinigung umfasst die Filterung um Dubletten, Semi-Dubletten sowie leere und

defekte Artikel. Der Wert, ab dem ein Artikelpaar als Semi-Dublette ausgezeichnet wird, kann ebenso wie Einstellungen zur Erfassung defekter und leerer Artikel vom Benutzer festgelegt werden. Die Anwender können sich Artikel anzeigen lassen, die die Zielwerte nicht erfüllen. Sie prüfen, ob es sich um defekte oder leere Artikel handelt, passen die Zielwerte entsprechend an und generalisieren anschließend auf das Sample. Die Artikel im Repository werden um Prozessmetadaten bereichert. Die Prozessmetadaten halten die Verarbeitungsschritte fest, die ein Dokument zu einem Zeitpunkt bereits durchlaufen hat.

2) *Complex-Concept-Builder*

Eine zentrale Rolle kommt einer Funktion zur Suche im Repository und der Darstellung von Ergebnissen zu. Wir stehen vor der Herausforderung, dass gesellschafts- und geisteswissenschaftliche Fragestellungen mit abstrakten Konzepten operieren, die sich nicht direkt in der Alltagssprache der aktuell untersuchten Zeitungstexte äußern. Kollektive Identitäten können ganz unterschiedlich ausgedrückt werden: „*wir Europäer* sind wir verpflichtet, den Völkermord im Land X zu stoppen“ ist ein sehr einfacher Fall. Typischer sind Ausdrücke wie „Deutschland sollte endlich Farbe bekennen“ oder „Washington kann in dieser Frage nicht über seinen Schatten springen“. Solche Appelle an unterschiedliche *kollektive Identitäten* verschiedener *politischer Akteure* sind zudem in den Zeitungsberichten verhältnismäßig selten. Um diese Instanzen im Meer der Worte leichter identifizieren zu können, entwickeln wir den *Complex-Concept-Builder*. Er umfasst Werkzeuge zur Topic-Analyse für die Bereinigung um Off-Topic-Artikel und für die Erstellung von Sub-Samples potentiell inhaltlich relevanter Artikel. Wir integrieren dazu über das Mallet-Tool eine Methode zur Textklassifikation, basierend auf Latent Dirichlet Allocation. Die Anwender wählen eine Anzahl n zu differenzierender Topics für ihr Sample. Die Werkzeuge klassifizieren die Texte in n thematisch definierte Gruppen, die den Anwendern präsentiert werden. Die Anwender können nun anhand dieser Beispiele bestätigen oder verneinen, ob die Artikel für ihre Fragestellung relevant sind oder nicht. Auf dieser Basis ergibt sich eine Wahrscheinlichkeit für jeden Artikel zum gesuchten thematischen Bereich zu gehören.

Schließlich haben wir ein Werkzeug zur Berechnung und Visualisierung der Medienaufmerksamkeit im Zeitverlauf („Issue Cycles“) in den Medien behandelter Themen entwickelt. Anwender können hier Themen und zu beobachtende Zeiträume definieren und in einer Zeitreihe darstellen. Mit dem Tool kann die Medienaufmerksamkeit für verschiedene Themen angezeigt werden und Perioden höherer Aufmerksamkeit für ein Thema können identifiziert werden.

Über die *Explorationswerkbank* werden sprachtechnologische Werkzeuge der CLARIN-Infrastruktur als Webservices eingebunden. Sie umfassen Wortartenerkennung, syntaktisches Parsing, Named Entity Recognition, Koreferenzanalyse und Sentimentanalyse. Diese Werkzeuge bilden Bausteine für die Erfassung komplexer Konzepte und können rechenintensiv sein. Wir wenden daher hier das Prinzip der modularen Ablaufketten an: Die Anwender wählen ein verfügbares Analysewerkzeug oder ein angestrebtes Ereignis aus, woraufhin die *Explorationswerkbank* die benötigten sprachtechnologischen Verarbeitungsschritte nennt. Die durchgeföhrten sprachtechnologischen Analyse-schritte werden als Prozessmetadaten zusammen mit den bearbeiteten Texten im Repository festgehalten.

Unsere Werkzeuge eröffnen weitreichende Möglichkeiten zur Exploration digitaler Textdaten. *Explorationswerkbank* und *Complex-Concept-Builder* versetzen Wissenschaftler aus den Digital Humanities zudem in die Lage, ohne die Mitarbeit von Tool-Experten die Potentiale der sprachtechnologischen Analyse ihrer Daten auszuprobieren und anzuwenden. Damit weisen sie Wege zu einer noch stärkeren Verbreitung sprachtechnologischer Methoden in den unterschiedlichsten mit großen Textmengen arbeitenden Fächern.

Quo Vadis Musikphilologie?

Digitale Ausgaben im Gespräch

Seit nunmehr etwa zehn Jahren wird im Kontext musikwissenschaftlicher Gesamtausgaben an der Entwicklung von Konzepten und der praktischen Umsetzung computerbasierter Editionsverfahren gearbeitet. Das Potential dieser Entwicklungen wurde im Fach bereits sehr früh erkannt, und trotz des häufig attestierten „konservativen Grundzugs“ der Musikwissenschaft zeichnet sich bereits heute ein grundsätzlicher Umbruch hin zu digitalen Arbeitsweisen ab. Vor allem die neu begonnenen Editionsprojekte der letzten Jahre sehen digitale Komponenten vor, aber auch bestehende Ausgaben werden zunehmend in diese Richtung weiterentwickelt. Die jeweiligen Konzepte weisen dabei eine große Bandbreite auf und reichen von der Bereitstellung erweiterter Materialien aus dem Kontext der weiterhin traditionell erscheinenden Ausgaben über in unterschiedlichem Maße aufbereitete Faksimiles bis hin zu vollständig digital erscheinenden Editionen, die bewusst mit den medial bedingten Konventionen der letzten 150 Jahre brechen, um neue Möglichkeiten zu erproben. Vor allem hinsichtlich der Positionierung zu herkömmlichen Editionsverfahren gibt es daher massive Unterschiede. Das Spektrum reicht dabei von weitgehend losgelösten Zusatzangeboten und reinen Retrodigitalisierungen, die lediglich um mediengerechtere Erschließungsmöglichkeiten erweitert werden, weiter über weitgehend traditionelle Buchpublikationen mit digitaler Beilage, welche Zugriff auf die verwendeten Quellenmaterialien bietet, bis hin zu rein digitalen Projekten, welche per se keinen unmittelbar in der Praxis verwertbaren Notentext mehr bereitstellen. Diese Vielfalt allein nach dem Kriterium einer scheinbaren „Modernität“ zu beurteilen, wäre eindeutig zu kurz gegriffen. Für einen qualifizierten Vergleich müssen vielmehr möglichst viele Parameter digitaler Ausgaben beleuchtet werden:

- Was ist das Ziel der Ausgabe, und an welche Zielgruppe richtet sie sich damit?
- Wo verortet sie sich im Spannungsfeld zwischen Wissenschaft und musikalischer Praxis?
- Welche Ressourcen stehen für die digitalen Bestandteile zur Verfügung, und in welchem Verhältnis stehen sie zur inhaltlichen Arbeit?
- In welcher Weise werden Faksimiles und / oder Codierungen genutzt, und wie sieht ggf. eine Aufgabenteilung zwischen diesen Komponenten aus?
- Welche Rolle wird dem Konsumenten der Ausgabe zugedacht – die des reinen Rezipienten vorab klar definierter Erkenntnisse, oder die des aktiven Benutzers, der eigene Fragen an das edierte Material stellen darf und soll?
- Wie wird mit einer möglicherweise veränderten Rolle des Editors umgegangen? Mit welchen Herausforderungen sieht sich ein Editor im digitalen Umfeld konfrontiert? Gibt es neue Arbeitsformen? Wie wird mit den steigenden Informationsbedürfnissen umgegangen?
- Besteht eine Zusammenarbeit mit Musikverlagen? Welche Rolle übernehmen diese, gerade auch im Bezug auf die digitalen Aspekte? Was wäre hier das Idealbild?
- In welcher Weise werden Informationen aus dem Kontext der Edition, etwa Briefe, Rezensionen oder Tagebucheinträge, mit einbezogen?
- Wird der Aspekt der Langzeitarchivierung im Projekt aktiv thematisiert? Welche Lösungen wurden in Bezug auf Formate und Zuständigkeiten für einen dauerhaften Zugang gefunden?
- Inwiefern ergeben sich durch die digitale Arbeitsweise neue rechtliche Herausforderungen, sowohl im Hinblick auf die urheberrechtliche Verwertung der Ausgabe als auch auf die Lizenzierung von digitalisiertem Quellenmaterial?

Anhand dieser Fragestellungen sollen einige zentrale Projekte im Bereich der digitalen Musikedition vorgestellt und ihre jeweiligen Konzepte im Vergleich diskutiert werden. Die beiden letztgenannten Bereiche – Langzeitarchivierung und Urheberrecht – werden dabei bewusst ausgeklammert, da deren Thematisierung

Quo Vadis Musikphilologie?

angesichts des Umfangs der Problematik zweifellos je eigene Panels erfordern würde. Vielmehr soll der Blick auf die methodischen Veränderungen der editorischen Arbeit selbst gerichtet werden, deren Tragweite oft genug durch die vorrangige Diskussion der technischen und rechtlichen Konsequenzen unbeachtet bleibt. Im Anschluss an kurze Eingangsstatements der Diskussionsteilnehmer (mit einer knappen Vorstellung des jeweiligen Projekts bzw. des Projektkontexts) sollen diese methodischen Aspekte zunächst auf dem Podium diskutiert werden, die Diskussion soll dann aber für das Plenum geöffnet werden. Die nachfolgend vorgestellten Projekte werden sich am Panel beteiligen:

www.schubert-online.at ist eine Online-Datenbank, welche digitale Reproduktionen von gegenwärtig mehr als 500 Notenautographen, Briefen und Lebensdokumenten Franz Schuberts enthält. Damit stellt sie die umfangreichste Sammlung von Schubert-Autographen im Internet dar. Der Grundgedanke dieser Datenbank ist es, die wissenschaftliche Arbeit mit Handschriften zu unterstützen, indem die Bedingungen und Voraussetzungen dafür am Computer simuliert werden. Dadurch können die Autographen einerseits geschützt und ihre Erforschung andererseits beschleunigt werden. Zu sämtlichen Materialien finden sich ausführliche Quellenbeschreibungen, für alle Textzeugen werden zudem Übertragungen angeboten. Eine überaus hilfreiche Besonderheit ist die virtuelle Zusammenführung von in verschiedenen Bibliotheken lagernden Handschriften-Fragmenten, die nur so vollständig erfasst werden können. www.schubert-online.at stellt damit bereits seit 2006 eine unverzichtbare Ressource für eine wissenschaftliche Auseinandersetzung mit den Originalquellen dar, bietet per se aber keine Edition im eigentlichen Sinne.

Im Rahmen der 2011 gestarteten *Kritischen Ausgabe der Werke von Richard Strauss* entsteht eine Sammlung von TEI- / MEI-codierten Briefen und Rezeptionszeugnissen, die in Form von werkbegleitenden Online-Dokumentensammlungen publiziert werden sollen. Während Möglichkeiten zur datenbankgestützten Verwaltung und Bearbeitung dieser Dokumente bereits eingerichtet sind, werden gegenwärtig technische Hilfsmittel zur Gruppierung und auszugsweisen Zitierung von Dokumentinhalten entwickelt. Dadurch soll es den Editoren ermöglicht werden, Inhalte zentral vorzuhalten und ggf. zu korrigieren, sie aber ohne Mehrarbeit in verschiedenen Kontexten in dieser je aktuellen Form zu nutzen. In dieser frühen Phase richtet sich der Blick des Projekts damit zunächst auf die Arbeit des Editors, wobei zukünftige weitere Nutzungsmöglichkeiten der so aufbereiteten Daten ausdrücklich nicht ausgeschlossen werden.

Die *Digitale Mozart Edition* stellt mit über 25.000 Seiten Notentext zweifellos die gegenwärtig umfangreichste digitale musikwissenschaftliche Ausgabe dar. Vorläufig handelt es sich dabei um Retrodigitalisate der seit 1954 erschienenen *Neuen Mozart Ausgabe*, die vollständig, d.h. inklusive Kritischer Berichte im Internet zur Verfügung stehen. Als besonders hilfreich erweist sich dabei die Parallelisierung von Ediertem Text und Kritischem Bericht, wodurch sich die Inhalte der Edition erheblich leichter erschliessen lassen als im gedruckten Band. Intern werden derzeit weitere Ausbaustufen der *Digitalen Mozart Edition* erarbeitet, die sich vom aktuell eher statischen Konzept deutlich lösen und über eine Codierung sämtlicher Notentexte eine Prozessierbarkeit der eigentlichen Editionsinhalte herstellen werden.

Die *Reger-Werkausgabe (RWA)* – begonnen 2008 mit der I. Abteilung »Orgelwerke« – verbindet als hybrid angelegtes Editionsprojekt konventionell gedruckte Notenbände mit digitalen Beigaben auf DVD. Diese Zusatzinhalte bilden für den Benutzer aufgrund der Materialfülle (Quellenabbildungen, umfangreicher lexikalischer Teil) nicht nur einen entscheidenden Mehrwert, den zu liefern in einem gedruckten Band schlicht unmöglich wäre, sondern sind zugleich ein essentieller Bestandteil der Edition: So ist z.B. der vollständige Kritische Bericht samt Lesartenverzeichnis (also der philologisch notwendige Apparat einer wissenschaftlich-kritischen Edition) nur auf der beigefügten DVD enthalten, im Druckband erfolgt dagegen eine Konzentration auf diejenigen Bemerkungen, welche die klangliche Werkgestalt betreffen, also vor allem für den Interpreten von Interesse sind. Beide Bestandteile der Edition sind je nach Benutzerinteresse unabhängig voneinander verwendbar.

Das Projekt *OPERA* beabsichtigt, herausragende Werke des europäischen Musiktheaters in exemplarischen Einzelausgaben vorzulegen. Die Auswahl der Kompositionen erfolgt dabei zum einen nach der musiktheatergeschichtlichen Bedeutung des jeweiligen Werkes, die sich vor allem aus dessen kompositions- und gat-

Quo Vadis Musikphilologie?

tungsgeschichtlichem Rang ergibt, zum anderen aus der mit der jeweiligen Komposition verbundenen edito□rischen Problemstellung. Der wesentliche Unterschied zur *RWA* ist damit die Abkehr von der autorbezogenen Perspektive üblicher Gesamtausgaben. Aus methodischer wie auch technischer Sicht bedeutsam ist die enge Verzahnung von Libretto- und Musikedition. Wie die *RWA* erscheint *OPERA* als traditionelle Print-Edition mit digitaler Beilage.

Im Projekt *Freischütz Digital* (*FreiDi*) soll anhand eines an Frans Wierings *Multidimensional Model* ange□lehnten Konzepts genuin digitaler Musikditionen am Beispiel von Carl Maria von Webers *Freischütz* ein proof of concept sowohl für die Möglichkeiten neuartiger Editionsmethoden als auch damit verbundener neuer Fragestellungen geliefert werden. Bezeichnend dafür ist die Abkehr vom Konzept *eines* Edierten Tex□tes: Die im digitalen Medium verzichtbare Festlegung auf einen in erster Linie aufführungspraktisch eingedruckten Werktext wird hier durch das Aufzeigen von Alternativen ersetzt, die einerseits die oft vorhandene Mehrdeutigkeit etwa im Bereich der akzidentellen Partiturbestandteile besser abzubilden vermag, andererseits dem Benutzer der Ausgabe einen weit größeren Einblick in die hinter der Ausgabe stehende editorische Arbeit erlaubt. Über eine detaillierte Codierung der musikalischen Quellen, aber auch der Textvorlage und anderer Materialien, wird dabei bewusst versucht, die Machbarkeit alternativer Editionskonzepte auszuloten.

Für das Panel haben sich die folgenden Kolleginnen und Kollegen zur Verfügung gestellt:

- Dr. Walburga Litschauer, *Schubert Online*
- Dr. Stefanie Steiner-Grage, *Reger-Werkausgabe*
- Dr. Alexander Erhard, *Richard Strauss-Ausgabe*
- Mag. Franz Kelnreiter, *Digitale Mozart-Edition*
- Dr. Andreas Münzmay, *OPERA*
- Benjamin Wolff Bohl M.A., *Freischütz Digital*

Aufgrund etwaiger terminlicher Schwierigkeiten ist es möglich, dass sich diese Liste der die Projekte jeweils vertretenden Personen bis zur Tagung noch ändern wird, in jedem Fall ist aber für eine Vertretung aus dem jeweiligen Projekt gesorgt, so dass sich an der grundsätzlichen Zusammenstellung nichts ändern wird. Die Moderation des Panels übernimmt der Einreichende.

Johannes Kepper

Semantisch-Soziale Netzwerkanalyse am Beispiel buddhistischer Texte in der Pali-Sprache:

Zwischenstand zur Korpus-Aufbereitung

von Jürgen Knauth und Sven Wortmann

Projekt: SeNeReKo (Uni Bochum/ Uni Trier)

Ziel des Projekts „Semantisch-Soziale Netzwerkanalyse als Instrument zur Erforschung von Relgionskontakten“ (SeNeReKo) ist es maschinelle Analyseverfahren auf antike religiöse Textkorpora anzuwenden um herauszufinden, welche semantisch zentralen Begriffe mit welchen religiösen Akteuren verknüpft werden sowie die narrativen Muster von interreligiösen „othering“-Prozessen aufzuzeigen. SeNeReKo besteht aus zwei Teilprojekten zu altägyptischen Texten und buddhistischen Pali-Texten.

Im Teilprojekt zum buddhistischen Pali-Kanon werden große Textmengen in der mittelindischen Sprache Pali - einer indoeuropäischen Sprache ähnlich dem Latein und Altgriechischen - analysiert. Der buddhistische Pali-Kanon wurde etwa um die Zeitenwende auf Sri Lanka zusammengestellt, enthält die heiligen Texte des Theravada-Buddhismus (verbreitet in Sri Lanka, Thailand, Burma, Vietnam) und ist von großem Wert für die Rekonstruktion des religiösen Feldes des antiken Indien. In den Narrativen des Pali-Kanons finden sich unzählige Belehrungs- und Bekehrungsepisoden zwischen dem Buddha und buddhistischen Mönchen auf der einen Seite sowie religiösen Konkurrenten (brahmanische Priester und andere Asketen) auf der anderen Seite. Eine maschinelle semantisch-soziale Netzwerkanalyse dieser Texte wird genauer als herkömmliche manuelle Analysen die umkämpften Begriffsfelder und sozialen Konstellationen dieser Texte darstellen und somit den diskursiven Startpunkt einer der größten Religionen der Welt beleuchten. Ein Nebenertrag könnte darin bestehen, anhand von Wortfeld-, Sprachmuster- und Stilanalysen Spuren von innerkanonischem Texttransfer zu finden und somit die nach wie vor weitgehend ungeklärte Kompositionsgeschichte des Pali-Kanons zu erhellen.

Ausgehend von der Annahme, dass für die von uns geplante Überführung der Textkorpora (bzw. von Teilen der Textkorpora) in semantische Netze die genauen Kenntnisse über die Wortart der einzelnen Wörter von großer Bedeutung sind, konzentriert sich der erste Teil der Arbeit in SeNeReKo auf eine geeignete Datenaufbereitung. Um eine Prozessierung der Texte zu einem späteren Zeitpunkt im Projekt zu erleichtern, wurde daher ein möglichst einheitliches Tagset festgelegt, welches nicht nur jeweils eine, sondern beide Textkorpora abdeckt. Aus Gründen der Erleichterung einer späteren Verwertbarkeit unserer Daten durch andere Wissenschaftler wurde bei der Erstellung dieses Tagsets auf Standards Wert gelegt: Alle Tags verweisen daher auf Einträge in der ISOCat-Datenbank. Um die Texte mit geeigneten PoS-Tags auszuzeichnen, mussten jedoch textspezifische Wege beschritten werden, die sich im Altägyptischen und im Pali auf Grund der völlig unterschiedlichen Datenlage entsprechend voneinander unterscheiden. Das Ergebnis dieser Verarbeitung sind Texte im TEI-Format. Diese sind für Visualisierungen und weitere Verarbeitungsschritte über

einen Konverter in andere Datenformate überführt worden und liegen daher neben TEI ebenso in TCF-Format vor, sowie in einer HTML-Repräsentation.

Die Ausgangsdaten des Pali-Kanons waren eine Gruppe von etwa 2700 Einzeldateien in rudimentärer TEI-Auszeichnung vor. Der Strukturierungsgrad dieser Daten war nur wenig höher wie Fließtext. Daher wurde zu Beginn des Projektes eine Segmentierung und Tokenisierung der einzelnen Texte des Korpus Sätze, Wörter – und soweit vorkommend – sonstige Daten vorgenommen und in einem geeigneten TEI-Format gespeichert. Fast alle im Ausgangsmaterial vorhandenen Informationen wurden dabei beibehalten, so dass dieser Schritt im Grunde eine invertierbare Transformation darstellte.

In der weiteren Aufbereitung der Daten sollen die in den TEI-Daten enthaltenen Wörter mittels automatisiertem PoS-Tagging ausgezeichnet werden. Auf Grund des im Vergleich zu z.B. Latein geringeren Reichtums des Pali an Wortendungen ist ein statistisches Modell als Grundlage für ein PoS-Tagging unumgänglich. Daher wurde in Hinblick auf die Erzeugung eines solchen statistischen Modells 1000 Sätze nach dem Zufallsprinzip aus dem Korpus extrahiert, die manuell getaggt wurden (und weiterhin getaggt werden). Ein größerer Teil dieser Sätze liegt nun bereits in vollständig annotierter Form vor und kann für die wissenschaftliche Arbeit verwendet werden.

Der Auszeichnungsprozess mit Part-of-Speech-Tags ist im vorliegenden Fall einigen Besonderheiten unterworfen: Zum einen handelt es sich beim Pali um eine historische, tote Sprache. Die Auszeichnung muss daher durch nicht-Native-Speaker geschehen. Zum anderen finden sich in Pali Sandhis (Laufveränderungen), welche im vorliegenden Tagging-Prozess gesondert beachtet werden: Diese werden während des Taggens manuell aufgespalten. Geeignete Daten für die zukünftige Verarbeitung müssen Sandhis in bereits aufgelöster Form enthalten, daher ist bereits in unserem PoS-Tagging diese Aufspaltung vorgesehen. Ein Tagging-Werkzeug, welches für diesen Arbeitsschritt verwendet wird, müsste dies unterstützen. Da abgesehen davon neben einem besonderen Augenmerk auf gute Usability ferner Hilfswerzeuge benötigt werden und werden, die das Taggen von Teilen des Korpus vereinfachen, wurde die Entwicklung eines eigenen Tagging-Tools initiiert. Aus Gründen der Benutzerfreundlichkeit und Realisierungseffizienz basiert dieses Werkzeug nicht auf Web-Technologien, sondern ist als Desktop-Applikation angelegt. Es speichert jedoch analog zu Web-Anwendungen alle relevanten Daten auf einem Server und kann damit Anforderungen erfüllen, die in der Regel sonst nur von Web-Anwendungen erfüllt werden.

Das Tagging-Tool wurde auf Grund der begrenzten Projektkapazitäten auf die spezifischen Anforderungen im SeNeReKo-Projekt abgestimmt, ist aber grundsätzlich universell nutzbar angelegt. So beinhaltet es zwar auch projektspezifische Besonderheiten wie das Auflösen der oben erwähnten Sandhis, doch kann es mit auch in Zukunft für andere wissenschaftlichen Projekten genutzt werden. Vom effizienten User-Interface könnten gerade Projekte zu historischen Daten in Zukunft profitieren.

Die durch dieses Werkzeug erstellten Pali-Referenz-Daten stehen nun als Trainingskorpus für das Erstellen von Modellen für maschinelles PoS-Tagging zur Verfügung. Dieser Arbeitsschritt ist gegenwärtig „Work in Progress“.

Den Prozess des PoS-Taggings muss ein Prozess der Lemmatisierung des zu bearbeitenden Pali-Korpus begleiten, damit später semantische Netze korrekt erzeugt werden können: Wörter in flektierter Form sind in unserem Fall für eine Aufbereitung in solche Netze nicht nutzbar. Da eine solche Lemmatisierung eine umfassendes Datenbank aller Wortformen benötigt wird, wurde versucht, eine solche Datenbank zu realisieren. Diese kann prinzipiell aus den Einträgen eines regulären Pali-Dictionaries erzeugt werden.

Da es sich im Pali um keine gängige (lebende) Sprache handelt, gibt es leider jedoch kein vollständiges, computerlinguistisch nutzbares Wörterbuch des Pali. Daher konzentrierte sich ein Teil der bisherigen Arbeit auf die Aufbereitung eines bestehenden (gedruckten) Wörterbuchs der Pali Text Society, welches wir in rudimentärer digitaler Form von der Library of Chicago erhalten konnten. Die aufbereiteten Daten wurden dabei in einen eigens für das Projekt entwickelten Wörterbuchserver eingespeist, der dahingehend konzipiert und entwickelt wurde, dass unterschiedliche Personen und unterschiedliche Werkzeuge in Zukunft parallel und unabhängig voneinander mit einem und demselben Wörterbuch arbeiten können: Dem Aspekt der Verteiltheit und aktiven Arbeiten mit den Daten kommt hierbei besondere Bedeutung zu.

Die Realisierung dieses Wörterbuchservers erfolgte auf Grund eines möglichst schnellen Entwicklungszyklus und möglichst guter Performance durch Verwendung der NoSQL-Datenbank „MongoDB“ und einer davor gesetzten NodeJS-Webapplikation. Es handelt sich also dabei um eine klassische 2-Tier-Architektur. Hier treffen sich informatische Gesichtspunkte, insbesondere Aspekte der Softwarearchitektur mit Anforderungen der Digital Humanities: Durch diese Architektur wird nicht nur sicher gestellt, dass (wörterbuchserverspezifische) Applikationslogik nicht in unterschiedlichen Softwarewerkzeugen wiederholt implementiert werden muss, sondern gleichzeitig auch eine klar definierte (HTTP-basierte) Schnittstelle bereit gestellt, welche eine Unabhängigkeit darauf zugreifender Software von konkreten Programmiersprachen sicherstellt.

Der Server ist so gebaut, dass er Batch-Operationen unterstützt, um so netzwerkbedingte Latenzen bei den Zugriffen auf einzelne Wörterbucheinträge zu minimieren. Die Kombination dieses Konzepts mit den oben genannten Technologien erlaubt durchschnittliche Lese- und Schreiboperation von deutlich unter einer Millisekunde pro Wörterbucheintrag, so dass trotz des netzwerkbasierten Ansatzes und der dadurch normalerweise zu einem Problem auflaufenden Netzwerklatenzen es dennoch nicht zu performance-Problemen kommt, welche die Arbeit mit den Daten behindern könnten.

Zusammen mit einer an die Konzepte von Weblicht angelehnten Verwaltung einzelner softwaretechnischer Werkzeuge und einer XML-basierten Textdatenbank, die Textkorpus-Daten in TEI und TCF-Format liefern wird, steht in Kürze eine Infrastruktur zur Verfügung, welche eine saubere Datenverwaltung und Verfügbarhaltung für andere Projektteilnehmer gewährleistet. Gleichzeitig wird sie so den Aufwand für die nächsten Entwicklungsschritte reduziert und es so ermöglichen, dass die semantisch-soziale Netzwerkanalyse nicht nur in exemplarischen Beispielen oder fest vorgegebenen Teilen realisiert werden kann, sondern wahlfrei für beliebige Teile unserer historischen Textkorpora.

Poster: Relationen im Raum: Visualisierung topographischer Klein(st)strukturen - ein interdisziplinäres eHumanities Projekt

Ziel des Verbundvorhabens "RiR - Relationen im Raum" ist die Analyse und Visualisierung räumlicher Relationen zwischen Grabmalen jüdischer Friedhöfe aus neun Jahrhunderten. Anhand konkreter Forschungsfragen und im Austausch der beteiligten Kulturwissenschaftler und Informatiker wird ein "Topographie-Visualizer" entwickelt, der die Analyse unterschiedlichster Friedhofsensembles ermöglichen soll. Hierzu sollen Datenbankinformationen und zwar möglichst frei und forschungsbezogen konfigurierbar auf kleinteiligen Lageplänen visualisiert werden.

Die Zugänge zu Grabmalen, ihren Inschriften und ihrer Formensprache sind vielfältig. In traditionellen fachwissenschaftlichen Editionen erfolgen diese meist chronologisch - Inschrift für Inschrift. Die Genealogie extrahiert verwandschaftliche Beziehungen und erstellt Stammbäume und Familientafeln. Strukturierendes Element der kunstwissenschaftlichen Betrachtungen ist die äußere Gestalt des Grabmals.

Bei all diesen Zugängen geraten jedoch die räumlichen Bezüge der Einzelobjekte zueinander leicht aus dem Blick. Ihr räumliches Neben- und Hintereinander, die Reihen und Felder - das, was die Grabmale vor Ort verbindet und den Friedhof erst als Ensemble von Grabmalen konstituiert, der topographische Zugang wird verstellt. Selten stehen Grabmale in willkürlicher Ordnung. Wer neben wem zu ewiger Ruhe gebettet wird, unterliegt kaum je dem Zufall. Explizite, häufiger noch unausgesprochene Regeln und Muster definieren die räumliche Ordnung der Grabmale.

Diesen vielfältigen Relationen im Raum widmet sich seit August 2012 ein vom Bundesministerium für Bildung und Forschung im Rahmen der so genannten "eHumanities" Förderlinie gefördertes Verbundprojekt. Der Forschungsverbund vereint Partner aus verschiedenen kulturwissenschaftlichen Disziplinen – Judaistik/Jüdische Studien, Bau-, Architektur-, Kunst- und Geschichtswissenschaften, die sich in erfolgreich abgeschlossenen und laufenden Projekten der Erforschung und Erschließung zahlreicher historischer Friedhöfe widmen – mit ihren Partnern aus der Informatik, die über reiche Erfahrung mit Visualisierung und den digitalen Infrastrukturprojekten DARIAH-DE und TextGrid verfügen.

Visualisierung

Die Fläche eines Bildes hat bemerkenswerte Eigenschaften. Anders als ein Buch und ganz im Gegensatz zur Datenbank, kann man auf der Fläche eines Bildes Ideen sehr frei arrangieren. Ein wildes Denken kann hier stattfinden, das noch nicht von der Buchkultur in das Korsett der Logik, die Schritt für Schritt, Zeile für Zeile, Seite für Seite, vorgeht, gepresst wurde. Bilder, gerade solche der Kunst, hintertreiben eindeutige Blick-Ordnungen. Nimmt ein Bildteil auf ein anderes Bezug, kann man das oft nicht in präzise Worte fassen. Es geht darum Bildbezüge auch ohne logische Kategorisierung zuzulassen. Dies trifft gleichermaßen auch auf den Lage-

plan eines Friedhofs zu, auf dem sich Bezüge zwischen den Abbildungen der Steine, den Inschriften, mit einer Baubeschreibung, mit Situationsphotos und einer Fülle von im Idealfall in Datenbanken erfassten Einzelinformationen kombinieren und visualisieren lassen, ganz genau so, wie es die Problemlage und jeweilige Forschungsfrage erfordert.

Mustererkennung

Mit dem im Rahmen des Projektes zu erarbeitenden Topographie-Visualizers, dem Werkzeug, mit dem die Daten aus den jeweiligen Datenbanken mit einem Friedhofsplan interaktiv verknüpft werden, lassen sich bereits jetzt in drei Prototypen - durch verschiedene Abfrage- und Darstellungsmöglichkeiten mit wenigen Klicks Forschungsfragen entwickeln und beantworten, deren Bearbeitung früher viele Arbeitsstunden in Anspruch genommen hätte oder schlicht nicht zu bewältigen gewesen wäre.

So liest man nun plötzlich an den interaktiven Lageplänen Regelmäßigkeiten ab, etwa Reihen von Bestatteten, die alle das selbe Geschlecht haben. Was lässt sich nun aus einer Lücke in einer Reihe von bestatteten Männern schließen? Was kann man zusätzlich vermuten, wenn alle anderen Grabstellen chronologisch belegt wurden? So simpel die Hypothese auch klingt, so wenig wäre sie ohne eine Darstellung aller historischer Daten auf einem Plan aufstellbar gewesen: es wird sich wohl um die Stelle eines Mannes handeln, der zwischen den Bestattungsdaten der angrenzenden angereihten Männer verstarb.

Solcherart Fragen stellen und beantworten zu können, dient das Projekt.

Welche Erkenntnisse sich gewinnen lassen, wenn man die an verschiedenen Orten und mit sehr unterschiedlichen Methoden gepflegten Datenbestände aufeinander bezieht, das ist unsere Neugier.

Projektpartner

Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte, Essen
Institut für Kultur und Ästhetik digitaler Medien, Leuphana Universität Lüneburg
DAASI International GmbH, Tübingen
Bau- und Stadtbaugeschichte, Fakultät 6, Institut für Architektur, TU Berlin

Förderer

Bundesministerium für Bildung und Forschung

Vortrag: Relationen im Raum: Visualisierung topographischer Klein(st)strukturen - ein interdisziplinäres eHumanities Projekt

Ziel des Verbundvorhabens "RiR - Relationen im Raum" ist die Analyse und Visualisierung räumlicher Relationen zwischen Grabmalen jüdischer Friedhöfe aus neun Jahrhunderten (11.-20. Jhd.). Anhand konkreter Forschungsfragen und im Austausch der beteiligten Kulturwissenschaftler und Informatiker wird ein "Topographie-Visualizer" entwickelt, der die Analyse unterschiedlichster Friedhofsensembles ermöglichen soll. Hierzu sollen Datenbankinformationen und zwar möglichst frei und forschungsbezogen konfigurierbar auf kleinteiligen Lageplänen visualisiert werden.

Die Zugänge zu Grabmalen, ihren Inschriften und ihrer Formensprache sind vielfältig. In traditionellen fachwissenschaftlichen Editionen erfolgen diese meist chronologisch - Inschrift für Inschrift. Die Genealogie extrahiert verwandschaftliche Beziehungen und erstellt Stammbäume und Familientafeln. Strukturierendes Element der kunstwissenschaftlichen Betrachtungen ist die äußere Gestalt des Grabmals.

Bei all diesen Zugängen geraten jedoch die räumlichen Bezüge der Einzelobjekte zueinander leicht aus dem Blick. Ihr räumliches Neben- und Hintereinander, die Reihen und Felder - das, was die Grabmale vor Ort verbindet und den Friedhof erst als Ensemble von Grabmalen konstituiert, der topographische Zugang wird verstellt. Selten stehen Grabmale in willkürlicher Ordnung. Wer neben wem zu ewiger Ruhe gebettet wird, unterliegt kaum je dem Zufall. Explizite, häufiger noch unausgesprochene Regeln und Muster definieren die räumliche Ordnung der Grabmale.

Diesen vielfältigen Relationen im Raum widmet sich seit August 2012 ein vom Bundesministerium für Bildung und Forschung im Rahmen der so genannten "eHumanities" Förderlinie gefördertes Verbundprojekt. Der Forschungsverbund vereint Partner aus verschiedenen kulturwissenschaftlichen Disziplinen – Judaistik/Jüdische Studien, Bau-, Architektur-, Kunst- und Geschichtswissenschaften, die sich in erfolgreich abgeschlossenen und laufenden Projekten der Erforschung und Erschließung zahlreicher historischer Friedhöfe widmen – mit ihren Partnern aus der Informatik, die über reiche Erfahrung mit Visualisierung und den digitalen Infrastrukturprojekten DARIAH-DE und TextGrid verfügen.

Digitale Infrastruktur

Die zu visualisierenden Daten kommen aus verschiedenen Datenbanken. Dieselben Grabsteine werden beschrieben, aber aus verschiedenen disziplinären Perspektiven. Während die einen sich vorwiegend auf Personen und Inschriften konzentrieren, fokussieren sich die anderen auf die baugeschichtlichen Aspekte wie Material, Form. Daher die Notwendigkeit einer neuen Datenbank, die diese Daten zusammenführt und zusätzlich eine performante Suchmaschine zur Verfügung stellt.

Die Infrastruktur des Projektes wurde so aufgebaut, dass Datenänderungen in den Quelldatenbanken erkannt und in die neue Datenbank integriert werden und somit die Aktualität der Daten stets gewährleistet ist.

Perspektivenwechsel

Es ist erstaunlich, welche Wirkung ein Perspektivenwechsel haben kann. Alle Daten, die in diesem Projekt zu jüdischen Friedhöfen bearbeitet und dargestellt, sind bereits erhoben. Der veränderte Blick auf die Gesamtheit dessen, was vorliegt, verschafft den Betrachtern allerdings eine Schau, die alle diese Daten erst sinnvoll verbindet.

Visualisierung

Die Fläche eines Bildes hat bemerkenswerte Eigenschaften. Anders als ein Buch und ganz im Gegensatz zur Datenbank, kann man auf der Fläche eines Bildes Ideen sehr frei arrangieren. Ein wildes Denken kann hier stattfinden, das noch nicht von der Buchkultur in das Korsett der Logik, die Schritt für Schritt, Zeile für Zeile, Seite für Seite, vorgeht, gepresst wurde. Bilder, gerade solche der Kunst, hinterstreben eindeutige Blick-Ordnungen. Nimmt ein Bildteil auf ein anderes Bezug, kann man das oft nicht in präzise Worte fassen. Es geht darum Bildbezüge auch ohne logische Kategorisierung zuzulassen. Dies trifft gleichermaßen auch auf den Lageplan eines Friedhofs zu, auf dem sich Bezüge zwischen den Abbildungen der Steine, den Inschriften, mit einer Baubeschreibung, mit Situationsphotos und einer Fülle von im Idealfall in Datenbanken erfassten Einzelinformationen kombinieren und visualisieren lassen, ganz genau so, wie es die Problemlage und jeweilige Forschungsfrage erfordert.

Mustererkennung

Mit dem im Rahmen des Projektes zu erarbeitenden Topographie-Visualizers, dem Werkzeug, mit dem die Daten aus den jeweiligen Datenbanken mit einem Friedhofsplan interaktiv verknüpft werden, lassen sich bereits jetzt in drei Prototypen - durch verschiedene Abfrage- und Darstellungsmöglichkeiten mit wenigen Klicks Forschungsfragen entwickeln und beantworten, deren Bearbeitung früher viele Arbeitsstunden in Anspruch genommen hätte oder schlicht nicht zu bewältigen gewesen wäre.

So liest man nun plötzlich an den interaktiven Lageplänen Regelmäßigkeiten ab, etwa Reihen von Bestatteten, die alle das selbe Geschlecht haben. Was lässt sich nun aus einer Lücke in einer Reihe von bestatteten Männern schließen? Was kann man zusätzlich vermuten, wenn alle anderen Grabstellen chronologisch belegt wurden? So simpel die Hypothese auch klingt, so wenig wäre sie ohne eine Darstellung aller historischer Daten auf einem Plan aufstellbar gewesen: es wird sich wohl um die Stelle eines Mannes handeln, der zwischen den Bestattungsdaten der angrenzenden angereihten Männer verstarb.

Solcherart Fragen stellen und beantworten zu können, dient das Projekt.

Welche Erkenntnisse sich gewinnen lassen, wenn man die an verschiedenen Orten und mit sehr unterschiedlichen Methoden gepflegten Datenbestände aufeinander bezieht, das ist unsere Neugier.

Projektpartner

Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte, Essen
Institut für Kultur und Ästhetik digitaler Medien, Leuphana Universität Lüneburg
DAASI International GmbH, Tübingen
Bau- und Stadtbaugeschichte, Fakultät 6, Institut für Architektur, TU Berlin

Förderer

Bundesministerium für Bildung und Forschung

Alexander Koplenig

Prinzipielle Probleme der Anwendung statistischer Signifikanztests in der Korpuslinguistik

Wohl in kaum einem Bereich der Datenanalyse finden sich mehr Mißverständnisse, Fehlinterpretationen und Halbwahrheiten als bei der Anwendung und Interpretation von Signifikanztests, und zwar nicht nur bei Laien, sondern häufig auch bei gestandenen Fachleuten.
(Diekmann, 2002, S. 585f)

In dem Vortrag sollen anhand von quantitativen Beispielen und computergestützten Simulationen einige Argumente vorgestellt werden, die dafür sprechen, dass die Anwendungsvoraussetzungen, welche dem Prinzip des statistischen Signifikanztests zugrunde liegen, d.i. der Schluss von den Eigenschaften einer Stichprobe auf die Eigenschaften einer Grundgesamtheit, aus prinzipiellen Gründen in der Korpuslinguistik – als wichtige Teildisziplin der Digital Humanities – nicht erfüllt sind. Folgt man diesen Argumenten so ergeben sich für die Korpuslinguistik unter Umständen weitreichende Folgen.

In der Korpuslinguistik wird angenommen, dass die (relative Token-) Vorkommenshäufigkeit bestimmter sprachlicher Strukturen mit der kognitiven Repräsentation bzw. Prototypikalität dieser Strukturen verbunden ist, oder anders ausgedrückt, dass Korpushäufigkeit kognitive Verankerung instanziert. Findet man dann zum Beispiel heraus, dass gewisse sprachliche Strukturen in einem Korpus geschriebener Sprache häufiger auftreten als in einem Korpus gesprochener Sprache, so folgert man daraus, dass dieser Struktur ein wichtigerer Status im geschriebenen Diskurs verglichen mit dem gesprochenen Diskurs zukommt (Schmid, 2010). Das Ziel einer korpuslinguistischen Studie ist es also weniger Aussagen über die untersuchten Korpora zu tätigen als vielmehr von diesen Textsammlungen auf die sprachlichen Varietäten zu schließen, die sie als Ausschnitt repräsentieren sollen (Baroni & Evert, 2009, S. 2).

In diesem Zusammenhang wird typischerweise ein statistischer Signifikanztest verwendet, um zu belegen, wie sicher man sich sein kann, dass der gefundene Zusammenhang nicht nur zufällig aufgetreten ist oder auch wie sicher man sich sein kann, dass die Hypothese „in Wahrheit“ wirklich richtig ist. Diese Annahme trifft aus statistischer Sicht jedoch nur unter ganz bestimmten Voraussetzungen zu.

Ein kurzer Exkurs in die Bevölkerungswissenschaften soll dies näher beleuchten: Bei Wahlumfragen geht es zum Beispiel darum, mit Hilfe der Befragung einer Auswahl von wahlberechtigten Personen auf das Ergebnis der Wahl schließen zu können. Die Stichprobe besteht dabei aus den befragten Personen, während sich die Grundgesamtheit aus allen wahlberechtigten Personen zusammensetzt. Mit Hilfe der Daten lässt sich dann beispielsweise ein Zusammenhang zwischen Wahlabsicht und Beruf berechnen.

Entscheidend ist hierbei, dass die Personen, die befragt werden, per Zufall ausgewählt werden. Nur dann lassen es die Grenzwertsätze der Statistik zu, Eigenschaften der Grundgesamtheit über die Stichprobe zu quantifizieren (Jann, 2005, S. 124–127). Statistische Signifikanz hat dabei nichts mit der Wichtigkeit des Forschungsergebnisses zu tun. Vielmehr liefert ein statistischer Test „eine formale Entscheidungsregel, die aufgrund einer Stichprobe darüber entscheidet, ob [der gefundene Zusammenhang] für die Grundgesamtheit zutrifft.“ (Fahrmeir, Künstler, Pigeot, & Tutz, 2001, S. 404). Ein Signifikanzniveau von 0,01 beruht auf folgendem Gedankenexperiment: angenommen man würde die Zufallsauswahl und Befragung der wahlberechtigten Personen sehr häufig wiederholen, dann würde sich das in der tatsächlich vorhandenen Befragung erzielte Ergebnis nur in höchstens 1 Prozent aller (hypothetischen) Stichprobenziehungsinstanzen einstellen, obwohl es in Wirklichkeit, d.h. in der Grundgesamtheit, überhaupt nicht vorhanden ist.

Folgt man dieser Definition, so stellen sich für die Anwendung des Verfahrens in der Korpuslinguistik einige grundlegende Probleme: Ob sich ein Unterschied als statistisch signifikant erweist, hängt neben der absoluten Größe des Unterschieds vor allem von der Größe der Stichprobe ab. Das bedeutet, dass aufgrund ständig wachsender Korpusgrößen auch völlig unbedeutende Zusammenhänge signifikant werden.

Weiterhin gilt, dass man kein Korpus als eine (endliche) Zufallsstichprobe der ohnehin schwer definierbaren jeweiligen „Sprache als Ganzes“ im strikten statistischen Sinn bezeichnen kann (Baroni & Evert, 2009, S. 3). Findet man im eingangs erwähnten Beispiel einen signifikanten Unterschied zwischen dem Korpus geschriebener Sprache und dem Korpus gesprochener Sprache, so könnte bei Wahl einer anderen Korpusgrundlage auch das Ergebnis völlig anders ausfallen, was besonders bei seltenen sprachlichen Phänomenen problematisch sein kann.

Dabei bietet die statistische Methodologie keinerlei Hilfestellung bei der Beantwortung der Frage, welche von beiden Untersuchungen denn nun eher der „Wahrheit“ entspricht. Wenn jedoch verschiedene Untersuchungen mit unterschiedlichen Korpusgrundlagen in die gleiche Richtung

deuten, so kann man dies durchaus vorsichtig als Indikator für einen tatsächlich vorhandenen Unterschied zwischen den beiden sprachlichen Varietäten deuten.

Auf der anderen Seite wird dieses Problem zusätzlich dadurch verschärft, dass gewisse Arten von Texten (prinzipiell) nicht in einem Korpus erscheinen, man denke zum Beispiel an intime Gespräche zwischen Ehepartnern oder Diplomaten/-innen. Angenommen man würde solche Gespräche mit Einverständnis der Beteiligten aufnehmen, um sie anschließend in ein Korpus zu überführen, so ist davon auszugehen, dass die Aufnahme des Gesprächs reaktiv ist (Diekmann, 2002, S. 520–523), d.h. dass das eigentliche Gespräch durch die Messung beeinflusst wird. In Korpora geschriebener Sprache gilt dies in ähnlicher Weise für Werke, welche man aus Urheberrechtsgründen nicht veröffentlichen kann. Darüber hinaus werden ja gerade Zeitungstexte, die oftmals den Hauptbestandteil einer Textsammlung ausmachen, vor der Veröffentlichung nach bestimmten Regeln redigiert und können deshalb nur bedingt als prototypischer Ausschnitt der geschriebenen Sprache bezeichnet werden (Gries & Berez, noch nicht erschienen, S. 2).

Daher gilt, dass man ein Korpus als opportunistische Stichprobe der jeweils untersuchten Sprache bezeichnen muss (Lüdeling & Evert, 2005, S. 6). Gleichzeitig sei darauf hingewiesen, dass dies nur dann aus statistischer Sicht ein Problem ist, wenn die nicht vorhandenen Sprachbelege systematisch verzerrt sind, d.h., dass das untersuchte sprachliche Phänomen in den nicht vorhanden Texten anders repräsentiert ist (Diekmann, 2002, S. 357–359). Jedoch gibt es auch hier keine Methode, mit der man prüfen könnte, ob es sich um eine systematisch verzerrte Stichprobe handelt.

Nun könnte man einwenden, dass es in weiten Teilen der empirisch arbeitenden (sozial-)psychologischen Forschung gängige Praxis ist, die Testpersonen, die an einem Experiment teilnehmen aus einer studentischen Population zu rekrutieren. Da sich diese Gruppe ja durchaus systematisch von anderen Gruppen unterscheiden könnte – so die Argumentation weiter – sind auch hier die Voraussetzungen für Signifikanztests keineswegs erfüllt.

Dieser Einwand trifft jedoch aufgrund der inhärenten Forschungslogik eines Experiments nicht zu. Angenommen man führt ein psycholinguistisches Lesezeitexperiment in einem Forschungslabor durch. In dem Experiment möchte man zum Beispiel testen, ob ein mit Hilfe sprachwissenschaftlicher Kriterien erstellter verständlichkeitsoptimierter fachsprachlicher Text (vgl. Wolfer, Hansen, & Konieczny, 2013) schneller gelesen werden kann als der ursprüngliche Text.

Eine Studentin betritt das Labor und bekundet ihre Teilnahmeabsicht. Entscheidend ist nun, dass die Versuchsleiterin per Zufall entscheidet, ob die betreffende Person in die Experimentalgruppe (optimierter Text) oder die Kontrollgruppe (Ausgangstext) kommt.

Der Randomisation an dieser Stelle kommt eine zentrale Bedeutung zu, da sie das wohl fundamentalste Problem jedweder empirischer Untersuchung löst (Angrist & Pischke, 2008, S. 9–18). Zeigt sich ein Unterschied zwischen den beiden experimentellen Gruppen hinsichtlich der Lesezeit, so ist dieser Unterschied nicht darauf zurückführbar, dass es sich um eine Studentin handelt und diese eben schneller liest - die Studentin hätte ja ebenso gut in die eine wie in die andere Gruppen eingeteilt werden können. Der Effekt muss deshalb darauf zurückgeführt werden, dass es sich um den Einfluss der experimentellen Manipulation handelt, weil die beiden Gruppen ja bis auf Zufallsschwankungen völlig identisch sind. So ist zum Beispiel nicht davon auszugehen, dass in der Kontrollgruppe nur die langsam Leserinnen sind. Dies wäre zumindest bei einer halbwegs vernünftigen Stichprobengröße doch sehr unwahrscheinlich.

In der Korpuslinguistik könnte man als Lösung des Problems nun einfach dafür argumentieren, dass man sich mit Aussagen der folgenden Art begnügt: Es zeigt sich ein statistisch signifikanter Unterschied zwischen Korpus A und Korpus B. Dies hätte den Vorteil, dass man nicht davon ausgehen muss, dass die jeweiligen Korpora als Zufallsauswahl der jeweiligen sprachlichen Varietät fungieren und man nur beschreibt, was man in den jeweiligen Textsammlungen vorgefunden hat. In diesem Fall läuft jedoch ein statistischer Signifikanztest *per Definitionem* ins Leere. Zählt man zum Beispiel alle Instanzen eines bestimmten Modalverbs in Korpus A und vergleicht man diese mit dem Auftreten des gleichen Modalverbs in Korpus B, so hat man es überhaupt nicht mehr mit einer Stichprobe, sondern viel mehr einer Vollerhebung zu tun, weil man ja alle Elemente der jeweiligen Grundgesamtheit untersucht. Der inferente Schluss von der Stichprobe auf die Grundgesamtheit entfällt somit, weshalb sich Aussagen über die statistische Signifikanz eines Zusammenhangs eigentlich erübrigen.

Was folgt nun aus den hier angestellten Überlegungen?

Entscheidet man sich für eine rigorose Auslegung der Regeln der Inferenzstatistik, so gilt, dass statistische Signifikanztests eine mathematische Präzision vermitteln, die auf Grundlage von korpuslinguistischen Daten nicht haltbar ist. Aus diesem Grund müsste man strenggenommen generell auf die Verwendung von statistischen Signifikanztests verzichten. Andererseits gibt es auch gute Gründe (Diekmann, 2002, S. 600–601) die dafür sprechen die Berechnung von Signifikanztests zumindest als Orientierungshilfe für die Plausibilität eines Ergebnisses auch in korpuslinguistischen Untersuchungen beizubehalten.

Für welche der beiden Optionen man sich entscheidet, bleibt wohl letztlich Sache des individuellen Geschmacks. Ich hoffe jedoch in meinem Vortrag zu zeigen, dass Signifikanztests allein eine

sorgfältige inhaltliche Interpretation des erzielten Ergebnisses nicht ersetzen. Vielmehr sollten diese in jedem Fall durch Maße der Assoziations- bzw. Effektstärke ergänzt werden (Jann, 2005, S. 66–98), welchen unter Umständen sogar der Vorrang bei der Einordnung des Forschungsergebnisses gegeben werden sollte.

Literatur

- Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton NJ: Princeton University Press.
- Baroni, M., & Evert, S. (2009). Statistical methods for corpus exploitation. In A. Lüdeling & M. Kytö (Hrsg.), *Corpus linguistics: An international handbook* (Bd. 2, S. 777–802). Berlin: De Gruyter Mouton.
- Diekmann, A. (2002). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (8. Aufl.). Reinbek: Rowohlt Taschenbuch Verlag.
- Fahrmeir, L., Künstler, R., Pigeot, I., & Tutz, G. (2001). *Statistik: der Weg zur Datenanalyse ; mit 34 Tabellen*. Berlin [u.a.]: Springer.
- Gries, S. T., & Berez, A. L. (noch nicht erschienen). Linguistic annotation in/for corpus linguistics. In N. Ide & J. Pustejovsky (Hrsg.), *Handbook of Linguistic Annotation*. Berlin, New York: Springer. Abgerufen von http://www.linguistics.ucsb.edu/faculty/stgries/research/InProgr_STG_ALB_LingAnnotCorpLing_HbOfLingAnnot.pdf
- Jann, B. (2005). *Einführung in die Statistik*. München; Wien: Oldenbourg.
- Lüdeling, A., & Evert, S. (2005). The emergence of productive non-medical -itis. Corpus Evidence and qualitative analysis. In S. Kepser & M. Reis (Hrsg.), *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*. Berlin, New York: De Gruyter Mouton.
- Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (Hrsg.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* (S. 101–133). Berlin, New York: de Gruyter.
- Wolfer, S., Hansen, S., & Konieczny, L. (2013). *Are shorter sentences always easier? Discourse level processing consequences of reformulating texts*. Gehalten auf der European Society for Translation Studies 7, Gernersheim. Abgerufen von http://www.fb06.uni-mainz.de/est/Dateien/EST_2013_abstract_booklet_web.pdf

Wittgensteins Nachlass: „Semantische Annotation von Adjektiven im Big Typescript.“

Angela Krey

Centrum für Informations- und Sprachverarbeitung (CIS)

Ludwig-Maximilians-Universität München

angela.krey@campus.lmu.de

1. Einführung

Im vorliegenden Abstract soll die Posterpräsentation für semantische Annotation von Adjektiven im Big Typescript von Ludwig Wittgenstein vorgestellt werden. Die Annotation wurde im Rahmen des „Wittgenstein in Co-Text“ Projektes am Centrum für Informations- und Sprachverarbeitung (CIS) als Bachelorarbeit realisiert. Die Arbeit beschäftigt sich mit der Erstellung und Einführung semantischer Adjektivklassen in einem für das Big Typescript (Ts-213) erstellte Lexikon. Bezeichnet wird es als wiTTLex und ist im deutschen DELAF-Format gehalten. Es soll alphabetisch sortiert alle Wörter umfassen, die im gesamten Nachlass von Ludwig Wittgenstein vorkommen. Das Format ist speziell für die Arbeit mit lokalen Grammatiken und unserer Suchmaschine „wiTTFind“ (siehe <http://wittfind.cis.uni-muenchen.de>) gewählt. Formen eines Wortparadigmas, grammatische und syntaktische Informationen sind in einer bestimmten Kodierung enthalten. Dieses Lexikon ist an unsere Suchmaschine „wiTTFind“ gebunden und ermöglicht, dass mit der Suchmaschine nicht wie üblich nur nach Wörtern, sondern mithilfe der Klassifikation nach semantischen Wortklassen gesucht werden kann. An allen erfassten Stellen der durchsuchten Werke können

dadurch semantisch annotierte Adjektive gefunden, beziehungsweise entdeckt werden.

2. Klassifikation

Als Hauptkriterium für die Klassifikation dient nicht nur die Häufigkeit der Adjektive und deren Abdeckung, auch die Ansichten Wittgensteins zum Thema Klassifikation und Bedeutung sollen beachtet werden. Hierbei stellt sich schnell heraus, dass ein allgemeines Klassifikationsmodell für das Ts-213 nicht möglich ist und sogar eine autoren-spezifische Klasse, die Stilistika, erstellt werden muss. Außerdem spielen Farbadjektive eine große Rolle. Farbadjektive sind nicht nur die häufigsten Adjektive, Wittgenstein widmet Farben und Farbenmischung im Kapitel Phänomenologie des Ts-213 ein eigenes Unterkapitel. Aus diesem Grund wurde entschlossen, für Farbadjektive noch eine Unterklassifikation durchzuführen. Im Laufe der Klassifikation stellt sich immer wieder heraus, dass manche Wörter nicht eindeutig differenziert werden können und in mehreren Klassen vertreten sind. Zum Beispiel das Wort „lauter“ kann als Numeralia oder als die Steigerungsform von laut, die zur Klasse Komparativa gehört, auftreten. Das Wort „gut“ kann einer Eigenschaft einer Person entsprechen oder einer Evaluation.

Klasse	Kodierung	Kriterien	Beispiele
Farben	COL	Farben nach Farbenlehre Wittgensteins, Unterkategorie: Grundfarbe, Zwischenfarbe, Transparenz, Glanz, Farbigkeit	rot, klar, blau, gelb, schwarz, rein, klar
Numeralia	NUM	quantitativ	zwei, erst, einzig, viel, weniger, zweite
Relation	REL	„auf ein Wort verweisend oder sich darauf beziehend“/ Fachsprache/ „wie“/ „aus“	physikalisch, wesentlich, kausal, grammatisch, sinnlos
Eigennamen	EN	von Eigennamen abgeleitete Adjektive	Fermatschen, Skolemschen, Sheffersche, Eulerscher
Temporalia	TEMP	Zeitangaben	unmittelbar, ehe, lang, andermal, vorig, meist
Evaluation	EVAL	Wertung des Betrachters/ Konnotation/ Subjektivität	falsch, richtig, wahr, einfach, genau
Zustände	ZU	abgeschlossener/ definierbarer Zustand, Lokativa	ganz, wirklich, gegeben, gebraucht, gesagt, allgemein
Komparativa	KOMP	Ähnlichkeit/ Gleichheit/ Steigerung	anders, gleich, verschieden, analog, besser
Stilistika	STIL	autorenspezifische Abstufungen	wohl, wirklich, bestimmt, klar, gewiß
Eigenschaft	EIG	Eigenschaften die Personen zugeschrieben werden	gebildet, ungeschickt, peinlich, selbstständig, gutwillig, naiv
Ereignisse	ER	nicht abgeschlossener Verlauf/ Handlung	unendlich, hinweisend, entsprechend, vorbereitend, folgend

Abbildung 1: Semantische Klassifikation der Adjektive

3. Benutzerinterface

Um in unserer Suchmaschine „wiTTFind“ (siehe <http://wittfind.cis.uni-muenchen.de/semantik.php>) die neuen Suchoptionen benutzerfreundlicher zu gestalten, wurde für diesen Zweck eine „semantische Suche“ mit spezieller Eingabemaske für die neuen semantischen Klassen erstellt. Das Hauptkriterium soll die manuelle Auswahl der semantischen Kategorien sein. Die Eingabezeile der Suchmaschine wurde zu einem Eingabefeld umfunktioniert. Je nachdem, welche semantische Klasse mit einem Häkchen gekennzeichnet wird, erscheint diese in der Suchzeile. Des Weiteren sollen auch mehrere Kategorien miteinander kombiniert werden können, die Suche sich auf bestimmte Wortar-

ten beschränken, und auch den Kontext miteinbeziehen können. Mehrere Häkchen können gesetzt und somit mehrere semantischen Klassen kombiniert werden. Der Kontext der Suche kann zusätzlich durch Satzzeichen eingeschränkt werden, hierbei sind vor allem in Anführungszeichen gesetzte Wörter wie 'rot' oder „rot“ gemeint. Außerdem müssen die Kategorien durch Beispiele veranschaulicht werden. Diese Beispiele werden in Tooltips für jede Klasse gespeichert, die angezeigt werden, sobald per Mouseover der Tooltip aktiviert wird. Für einen noch besseren Überblick können bereits erstellte Frequenzlisten durch Anklicken der Beispiele dargestellt werden.

WiTTFind <ADJ+COL+NUM> WiTTFind-Suche

Suche einschränken

nur Adjektive <ADJ>

Semantische Klassen für Adjektive und Nomen

- Farben Beispiele für <ADJ> → anklicken für Frequenzliste
- Numeralia Beispiele für <ADJ>
- Eigennamen Beispiele für <ADJ>
- Temporalia Beispiele für <ADJ>
- Zustände Beispiele für <ADJ>
- Eigenschaft Beispiele für <ADJ>
- Ereignisse Beispiele für <ADJ>

Semantische Klassen nur für Adjektive

- Evaluation Beispiele für <ADJ>
- Relation Beispiele für <ADJ>
- Komparativa Beispiele für <ADJ>
- Stilistika Beispiele für <ADJ>

Abbildung 2: Beispielanfrage im Benutzerinterface

4. Evaluation

Die statistische Auswertung der semantischen Klassen und ihren Konkordanzen am Ende der Ar-

beit zeigt, dass die Adjektive jeder Klasse mit ähnlichen Nomen kombiniert werden können und so mit ein ähnliches Verhalten aufzeigen.

„Klöster und Stifte des Alten Reiches“ im Netz.

Ein Datenbankprojekt der Germania Sacra

Seit rund 100 Jahren erforscht die Germania Sacra die Kirche des Alten Reiches. Sie beschäftigt sich mit allen wichtigen Aspekten der Geschichte der Bistümer, Stifte und Klöster der Reichskirche. Als Grundlagenforschung ist es ihre Aufgabe, diese Informationen so bereitzustellen, dass sie für die weitergehende Forschung leicht zugänglich sind und möglichst vielfältig genutzt werden können. Hierzu gehört die Integration der durch sie erschlossenen Daten in das World Wide Web. Die Germania Sacra als traditionsreiche Forschungsinstitution bürgt für die Verlässlichkeit der Informationen.

Die Germania Sacra präsentiert ihre Ergebnisse nach wie vor als Handbücher im traditionellen Printformat. Mehr und mehr ist in den vergangenen Jahren die Notwendigkeit in den Fokus gerückt, die in den Printpublikationen veröffentlichten Informationen im Netz leichter, schneller und komfortabler zugänglich zu machen. Zum Kern der digitalen Angebote der Germania Sacra gehören die Digitalisate der Printpublikationen, die frei verfügbar im Internet angeboten werden. Eine tiefere Erschließung der wissenschaftlichen Informationen durch den Einsatz der technischen Möglichkeiten bietet das Digitale Personenregister der Germania Sacra, eine wissenschaftliche Personendatenbank mit derzeit über 20.000 Einträgen. Der inhaltliche Schwerpunkt liegt auf den Klerikern der Reichskirche.

In Ergänzung dieser Angebote wird die Germania Sacra im Frühjahr 2014 eine Online-Datenbank zu Klöstern und Stiften des Alten Reiches freischalten. Die wissenschaftliche Datenbank zielt darauf ab, ein Recherchetool zu schaffen, das regional übergreifend Basisinformationen zu allen Klöstern und Stiften auf dem Gebiet des Alten Reiches von der Zeit der Gründung monastischer Gemeinschaften bis zur Reformation bzw. Säkularisation bietet.

Die Datenbank soll die online verfügbaren wissenschaftlichen Informationen zu Klöstern und Stiften vernetzen. Hierfür stützt sich das Projekt auf die Arbeit mit Normdaten, Thesauri und Technologien des Semantic Web. Zudem wurde ein Datenmodell entwickelt, das die Zusammenarbeit mit ausgewählten Kooperationspartnern über den direkten Austausch von Daten leicht ermöglicht.

Im Internet werden bereits von einigen Bundesländern regionale wissenschaftliche Klosterdatenbanken angeboten. Die umfangreichen Detailinformationen, die von den Klosterprojekten der einzelnen Bundesländer bereitgestellt werden, werden durch den Nachweis der maßgeblichen verfügbaren Internetquelle in der Klosterdatenbank der Germania Sacra vernetzt.

In einer Reihe von Bundesländern wurden Klosterbücher herausgegeben, die zurzeit nur als Printversionen zur Verfügung stehen. Um auch die dort enthaltenen Basisinformationen verfügbar zu machen, werden diese in die Datenbank der Germania Sacra eingepflegt und mit einem Verweis auf den einschlägigen Klosterbuchartikel versehen.

Die Datenbank hält für alle Institutionen Basisinformationen bereit, die eine Recherche nach Ordenszugehörigkeit, zeitlichen Aspekten wie Gründung, Aufhebung und Dauer der Ordenszugehörigkeit und geographischer Lage ermöglichen.

Die Ergebnisse aller Recherchen sind in interaktiven Karten darstellbar, die die Klosterlandschaft des Mittelalters und der frühen Neuzeit visualisieren. Zeitschnitte und regionale sowie inhaltliche Aspekte sind dabei für den Nutzer frei wählbar.

Von der Klosterdatenbank aus hat der Benutzer direkten Zugriff auf die vom Forschungsprojekt herausgegebenen Monographien zu Klöstern und Stiften, die als Digitalisate frei verfügbar sind. Für alle verzeichneten Institutionen wird in der Datensatzanzeige das in den Germania-Sacra-Bänden erfasste geistliche Personal mit ausgegeben. Diese Personeneinträge sind direkt mit dem Digitalen Personenregister der Germania Sacra und den entsprechenden Fundstellen in den Online-Bänden verknüpft und ermöglichen so weiterreichende Recherchen.

Um die Vernetzung und Verdichtung der Informationen zu Klöstern und Stiften im World Wide Web zu fördern, bietet sich die Arbeit mit Normdaten an. Für viele der durch die Forschung der Germania Sacra generierten Informationen kann auf bereits vorhandene Normdaten zurückgegriffen werden. Besonders relevant für unser Projekt ist der Datenbestand der Deutschen Nationalbibliothek mit den dort verwendeten Datensatznummern der Gemeinsamen Normdatei (GND).

Für Personendaten wird üblicherweise das Beacon-Format verwendet, das das automatische Generieren von Links zu externen Datenquellen ermöglicht. Für andere Daten als Personen, etwa für Körperschaften, wird das Beacon-Format bisher kaum genutzt. Mit der Klosterdatenbank der Germania Sacra soll die Verwendung dieser Technik für Klöster und Stifte erprobt und eingeführt werden. Die Identifizierung der einzelnen Klöster und Stifte in der Gemeinsamen Normdatei der Deutschen Nationalbibliothek ist bereits vielfach erfolgt, fehlende Einträge in der GND werden durch die Germania Sacra ergänzt.

So können in der Datenbank automatisiert direkte Links nicht nur zu externen Klosterdatenbanken, sondern auch zu relevanten Datensätzen in Bibliothekskatalogen, Bestandsübersichten von Archiven, Quelleneditionen, Bibliographien, Porträtsammlungen und weiteren Informationsangeboten bereitgehalten werden.

Um den Möglichkeiten zur semantischen Recherche einen Weg zu bereiten, werden die Daten der Klosterdatenbank auf der Basis von Linked Data angereichert und im RDF-Format ausgegeben. Für die Klosterdatenbank wurde keine eigene Ontologie entwickelt, sondern es wird auf etablierte existierende Vokabulare zurückgegriffen. Für die Ausgabe der Datensätze im RDF-Format werden Normdaten für Orden, Bistümer, Personen wie auch Normdaten für Geografika (Geonames) verwendet. Vorhandene Einträge in der Wikipedia werden referenziert. Das modellierte Schema bietet hohes Potential, das Informationsnetz zu den Beziehungen von Personen und geistlichen Institutionen für den Zeitraum des Mittelalters und der Frühen Neuzeit zu verdichten.

1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014)

Universität Passau · 25.-28. März 2014

Thema 2: Digitale Infrastrukturen für die Geisteswissenschaften

Simone Kronenwett, Cologne Center for eHumanities, Universität zu Köln

Poster-Abstract

Bei fast allen Tagungen und Konferenzen zum Thema DH ist die Frage nach einer sinnvollen Weiterentwicklung der *digitalen Infrastrukturen für die Geisteswissenschaften* ein zentraler Diskussionspunkt.¹ Zwar besteht größtenteils Konsens darin, dass die Etablierung dedizierter digitaler Infrastrukturen für die Geisteswissenschaften im deutschsprachigen Raum derzeit führend ist. Auch wird der damit verknüpfte Professionalisierungsprozess der DH durch große laufende Infrastrukturprojekte wie Dariah und CLARIN weiter vorangetrieben und ausgebaut.² Gleichzeitig eröffnen sich damit weiterführende Fragen, die es zu beantworten gilt: Wie können stabile und damit langfristig finanzierte Infrastrukturen in den DH geschaffen und gesichert werden?³ Und wie können bestehende Desiderata beispielsweise hinsichtlich geisteswissenschaftlicher Forschungsdaten und Ressourcen geschlossen werden? Denn die Sicherung und die langfristige Zugänglichkeit von Forschungsdaten und Projektergebnissen ist auch ein maßgebliches Bewilligungskriterium aller Förderinstitutionen bei Projektanträgen im Sinne guter wissenschaftlicher Praxis.⁴ Die Wissenschaftsorganisationen haben deshalb in diesem Zusammenhang in den Strategiepapieren der letzten Jahre den Aufbau von entsprechenden Datenzentren empfohlen.⁵

Data Center for the Humanities

Um der konkreten Problematik einer dauerhaften Bereitstellung geisteswissenschaftlicher Forschungsdaten aktiv zu begegnen und damit die oben genannten Desiderata zu schließen, wurde Ende 2012 von der Philosophischen Fakultät der Universität zu Köln das Data Center for the Humanities (DCH) gegründet, dessen aktueller Entwicklungsstatus im Rahmen der Posterpräsentation vorgestellt wird.⁶

Das vom Cologne Center for eHumanities (CCeH) organisierte Datenzentrum zielt nicht nur darauf ab, die digitale Langzeitarchivierung von Forschungsdaten zu gewährleisten, sondern wird auch Forschungsdatenmanagement betreiben, dauerhafte Präsentationssysteme hosten und betreuen,

¹ So jüngst auf der *Closing Panel Discussion* der Herrenhausen-Konferenz der VolkswagenStiftung „(Digital) Humanities Revisited – Challenges and Opportunities in the Digital Age“, 5.-7. Dezember 2013, Hannover, <http://www.volksagenstiftung.de/digitalhumanities>.

² Vgl. Dariah-DE (Digital Research Infrastructure for the Arts and Humanities), <https://portal-de.dariah.eu/>; CLARIN-D (Common Language Resources and Technology Infrastructure), <http://de.clarin.eu/de/home>; vgl. auch BMBF (Hrsg.): Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften, Mülheim an der Ruhr, Stand: Februar 2013, S. 27ff., http://www.bmbf.de/pub/forschungsinfrastrukturen_geistes_und_sozialwissenschaften.pdf.

³ Als Vorbild wird immer wieder das naturwissenschaftliche Forschungsinfrastrukturprojekt CERN (Conseil Européen pour la Recherche Nucléaire, <http://home.web.cern.ch/>) genannt, vgl. BMBF: Forschungsinfrastrukturen, S. 7.

⁴ Vgl. Deutsche Forschungsgemeinschaft (DFG): Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“, Denkschrift, Weinheim 1998, http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf.

⁵ Vgl. Wissenschaftsrat: Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020, 13.07.2012, Berlin, S. 11, <http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf>.

⁶ Vgl. Data Center for the Humanities (DCH), www.dch.uni-koeln.de.

Services zur Datenkommunikation aufbauen und Werkzeuge zur Arbeit mit den Daten vorhalten.⁷ Die Zusammenarbeit und Kooperation mit verschiedenen Akteuren (wie z.B. der Universitäts- und Stadtbibliothek Köln oder dem Regionalen Rechenzentrum der Universität zu Köln) mit ihren jeweiligen Kompetenzen ist dabei für das Datenzentrum grundlegend. Im Mittelpunkt stehen allerdings die FachwissenschaftlerInnen und ihre Forschungsprojekte, deren Inhalte nachhaltig gesichert und dauerhaft zugänglich gemacht werden müssen.

Geisteswissenschaftliche Forschungsdaten

Unter Forschungsdaten werden gemeinhin sämtliche Daten verstanden, die als Grundlage für die Forschung dienen.⁸ Aufgrund der großen Methodenvielfalt in den geisteswissenschaftlichen Disziplinen zeichnen sich diese Daten besonders durch ihre Heterogenität aus.⁹ Um den Mehrwert der Daten in ihrer Kontextualität und Diskursivität zu wahren und nachnutzen zu können, kann häufig nicht einfach zwischen „Primärdaten“, die bloß zu archivieren wären, und den oft komplexen und individuellen Systemen der Präsentation und Benutzung von „Ergebnisdaten“ unterschieden werden. Neben der Langzeitarchivierung von Daten stehen deshalb die Pflege und Betreuung dieser komplexen Systeme im Fokus des Datenzentrums. Damit Forschungsdaten auch weiterhin vernetzt und kontextualisiert dargestellt und werden können, bedarf es eines Forschungsdatenmanagements und einer Begleitung der Forschung, die den Spezifika geisteswissenschaftlicher Fragestellungen und Methoden sowie der Präsentation ihrer Ergebnisse gerecht werden.

Geisteswissenschaftlicher Forschungsprozess

Geisteswissenschaftliche Forschung vollzieht sich in zyklischen Prozessen. In der Operationalisierung und Beantwortung von Forschungsfragen entstehen Daten, die selbst wieder das Ausgangsmaterial für neue Fragestellungen sein können. Damit Informationen dauerhaft nutzbar bleiben und Wissen nicht verloren geht, bedarf es eines professionellen Datenmanagements, das die Forschung begleitet und ihre Ergebnisse dokumentiert, mit Metadaten anreichert, archiviert, an Schnittstellen bereitstellt, präsentiert und für den andauernden Zugriff pflegt.

DCH-Schichtenmodell

Im Zentrum der DCH-Präsentation steht deshalb neben dem Lebenszyklus geisteswissenschaftlicher Forschungsdaten das Schichtenmodell des Datenzentrums. Archivierung, Bereitstellung, Adressierbarkeit, Präsentation und Nutzung von Diensten und Werkzeugen bauen hier aufeinander auf. Die modulare Struktur des Datenzentrums entspricht dabei den unterschiedlichen Anforderungen aus den verschiedenen Forschungsprojekten und gewährleistet insgesamt (1.) eine langfristige Sicherung und dauerhafte Bereitstellung, (2.) die allgemeine Zugänglichkeit und Nutzbarkeit und (3.) eine erhöhte Sichtbarkeit der Forschungsdaten und Ressourcen, die zugleich eine bessere Vernetzung der Projekte und Daten bedeutet und die Grundlage für die zukünftige Nutzung in Forschung und Lehre bildet.

⁷ Vgl. Cologne Center for eHumanities (CCeH), www.cceh.uni-koeln.de.

⁸ Vgl. z.B. Schwerpunktinitiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen: Grundsätze zum Umgang mit Forschungsdaten, 04. Juni 2010, <http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaezze/>.

⁹ Vgl. Jasmin Hügi, René Schneider: Digitale Forschungsinfrastrukturen in den Geistes- und Geschichtswissenschaften, Genf 2013, S. i, http://www.infoclio.ch/sites/default/files/standard_page/studie_forschungsinfrastrukturen_small.pdf.

Workshop

XML-Print

Ein Werkzeug zum Satz beliebiger XML-Dokumente

Prof. Dr. Marc W. Küster Lukas Georgieff Martin Sievers

16. Dezember 2013

1. Einleitung

XML ist unbestreitbar zu *dem* Dateiformat in den „Digital Humanities“ geworden. Insbesondere der TEI-Standard¹ wird von vielen neu entwickelten Werkzeugen als Ausgabe- und Austauschformat unterstützt. Auch ein XML-basierter Werkzeugkasten wie das TextGridLab², das die verschiedenen Entwicklungsstufen geisteswissenschaftlicher Projekte unterstützt, hat zur Verankerung von XML in der Anwendergemeinschaft maßgeblich beigetragen.

Bei aller Digitalisierung ist bei den Anwendern jedoch der Wunsch erhalten geblieben, ihre Texte bzw. Ergebnisse auch in eine gedruckte (analoge) Form zu überführen. Eine Umfrage der BBAW hat jüngst bestätigt, dass TextGrid-Anwender eine Printkomponente besonderes vermissen. Diese wird im Rahmen eines eigenständigen, DFG-geförderten Projekts *XML-Print* entwickelt und steht kurz vor der Fertigstellung. Unabhängig vom TextGridLab ist die Open-Source-Software allerdings auch eigenständig nutzbar und somit für einen großen Nutzerkreis von Bedeutung.

Nach einigen Live-Präsentationen und kleineren Workshops möchten wir auf der DHD 2014 nicht nur einen Einblick in die aktuelle Vorabfassung von XML-Print 2.0 geben, sondern allen interessierten Wissenschaftlern die Chance bieten, sich ausgehend von einer XML-Datei in einzelnen Schritten systematisch einer druckfertigen PDF-Ausgabe zu nähern.

2. Funktionsweise von XML-Print

2.1 Allgemein

XML-Print ermöglicht die Überführung beliebiger XML-Dateien in eine PDF-Datei.³ Grundlage dafür ist ein erweitertes Dateiformat XSL-FO+, das auf dem etablierten Standard XSL-FO⁴ beruht und diesen um bis dato fehlende typographische Elemente wie Mehrspaltigkeit und

¹ Vgl. Burnard und Bauman 2007.

² Mehr unter *TextGrid: Digital edieren – forschen – archivieren* 2013.

³ Andere Ausgabeformate können bei Bedarf über eigene Filter hinzugefügt werden.

⁴ Siehe dazu die Definitionen der *Extensible Stylesheet Language (XSL)* 2006, Version 1.1 sowie 2012, Version 2.0.

den Satz von Apparaten ergänzt. Die im Rahmen des Projekts entwickelte neuartige *Satzengine* interpretiert die XSL-FO+-Datei und erzeugt daraus eine druckfertige Ausgabedatei.

Obwohl man diesen Prozess komplett über die Kommandozeile steuern kann, ist es ein wesentliches Ziel von XML-Print, die Zuweisung von Layoutinformationen zu XML-Elementen über eine graphische Benutzeroberfläche komfortabel durchführen zu können. Der dazu entwickelte *Stileditor* erzeugt aus XML-Quelldatei(en) und im Hintergrund generierten XSLT-Stylesheets die für die Satzengine benötigte Eingabedatei (vgl. dazu auch Abbildung 1).

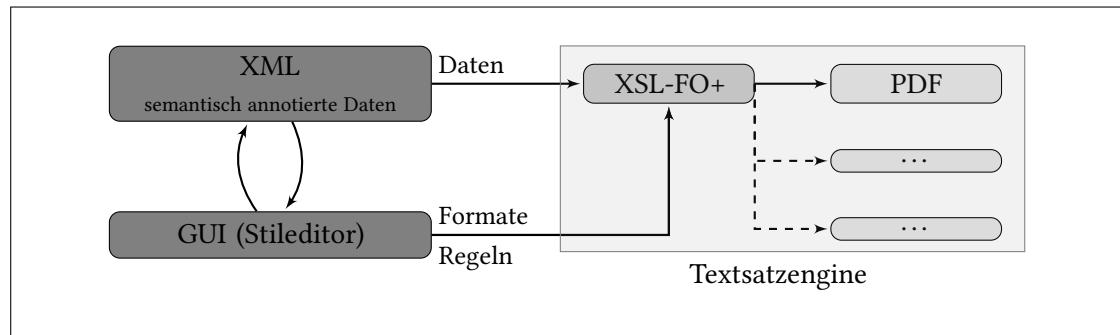


Abbildung 1: *Der Ablauf der Erstellung einer Druckausgabe mit XML-Print. Mit Hilfe der graphischen Oberfläche werden XML-Elementen Layoutinformationen zugewiesen und daraus das benötigte Zwischenformat generiert. Dieses wird von der Satzengine interpretiert und in eine PDF-Ausgabe überführt.*

2.2 Stileditor

Die graphische Benutzeroberfläche innerhalb von XML-Print, der *Stileditor*, verbindet die Quell-daten im XML-Format über XSLT-Templates mit Layoutinformationen und macht daraus eine Eingabedatei für die Satzengine im XSL-FO+-Format.

Der Anwender erstellt zunächst *Formate*, die das gewünschte Erscheinungsbild festlegen. Die verschiedenen Möglichkeiten sind zur besseren Übersichtlichkeit in Kategorien unterteilt (vgl. Abbildung 2). Anschließend definiert der Nutzer *Zuweisungen* zwischen einer Gruppe von XML-Elementen auf der einen und einem der Formate auf der anderen Seite. Die Auswahl der

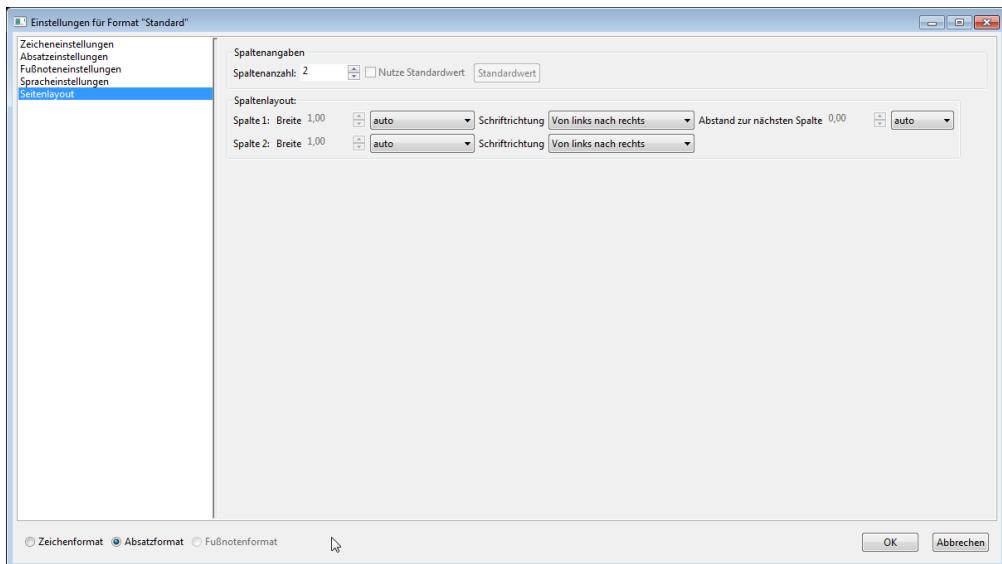


Abbildung 2: Formate beschreiben das gewünschte Aussehen bestimmter XML-Elemente. Zur besseren Übersicht sind die einzelnen Möglichkeiten in Kategorien unterteilt (linke Seite).

XML-Elemente geschieht dabei graphisch über ihre Position im XML-Baum oder alternativ über einen XPath⁵-Ausdruck (siehe Abbildung 3).

Über diesen Grundmechanismus lassen sich bereits sehr viele Layoutspezifikationen vornehmen. Für globale typographische Einstellungen wie Seitenformate oder die Definition von Feldern und Apparaten existieren zusätzlich jeweils eigene Dialoge (vgl. Abbildung 4).

2.3 Satzengine

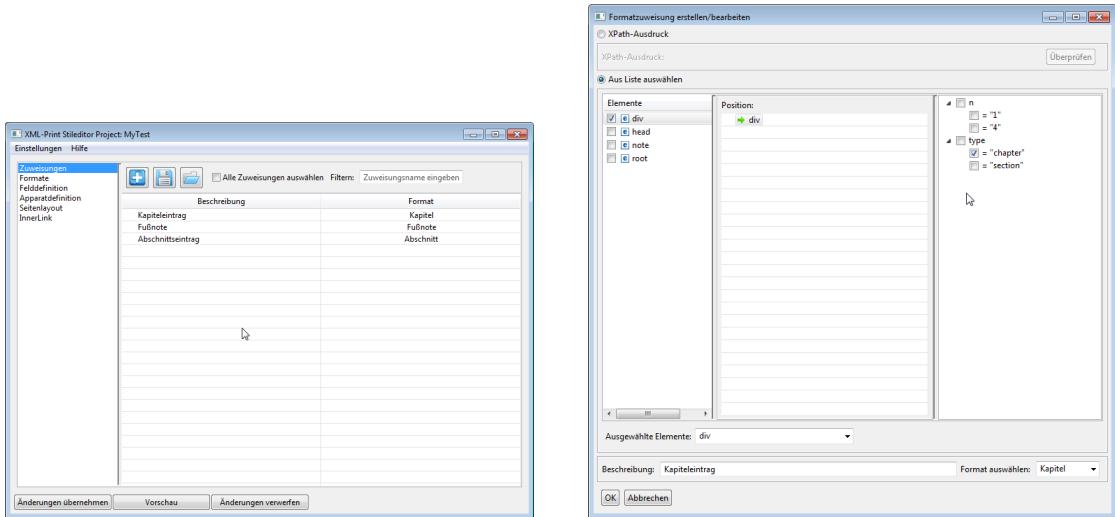
Es herrscht kein grundsätzlicher Mangel an Textsatzwerkzeugen mit hoher typographischer Ausgabequalität. Neben dem kommerziellen InDesign⁶ ermöglichen die Open-Source-Werkzeuge TeX⁷ sowie die Satzkomponente des TUSTEP⁸-Systems seit über dreißig Jahren hochwertigen Textsatz – allerdings erst, nachdem man sich in die Eigenheiten und -arten der Software eingearbeitet hat. Zudem sind beide Programme sehr Texteditor-lastig und damit für viele heutige

⁵ Die XPath-Syntax ermöglicht den Zugriff auf beliebige Teile eines XML-Dokuments. Zur Definition siehe *XML Path Language (XPath)* 2010.

⁶ Siehe <http://www.adobe.com/products/indesign.html>.

⁷ Zur Entstehungsgeschichte von TeX siehe z. B. <http://tug.org/whatis.html>.

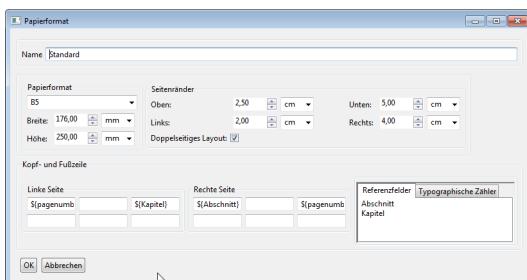
⁸ Siehe <http://www.tustep.uni-tuebingen.de/index.html>.



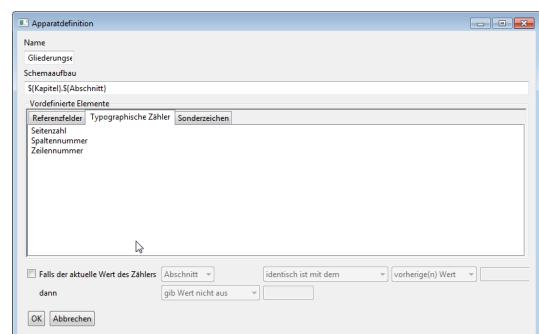
(a) Die Reihenfolge der Liste aller definierten Zuweisungen (rechte Seite) legt gleichzeitig die Priorität fest.

(b) Die Auswahl der XML-Elemente erfolgt wahlweise graphisch über ihre Position im XML-Baum oder über einen beliebigen XPath-Ausdruck.

Abbildung 3: Eine Zuweisung bestimmt für eine Gruppe von XML-Elementen, welches Format verwendet werden soll. Aus der Gesamtübersichts (a) gelangt man durch Auswahl einer Zuweisung zu den Einstellungen (b).



(a) Neben den Seiten- und Randmaßen werden für das Papierformat auch die Inhalte der Kopf- und Fußzeile definiert. Diese können in zwölf Bereichen aus benutzerdefinierten Feldern, typographischen Zählern und Zeichenketten beliebig gebildet werden.



(b) Für Apparateinträge können aus typographischen Zählern, benutzerdefinierten Feldern und Zeichenketten beliebige Referenzschemata gebildet werden. Zusätzliche Ausnahmen erlauben die selektive Ausgabe von Bestandteilen.

Abbildung 4: Globale Einstellungen in XML-Print. „Seitenlayout“ (a) und „Apparatdefinition“ (b) sind Kategorien der Hauptansicht (vgl. Abbildung 3a, linke Seite).

Anwender, insbesondere für Anfänger, die gewohnt sind, mit graphischen Systemen zu arbeiten, eher abschreckend.

Auch aus der Sicht der Informatik sind beide Programme – aus nachvollziehbaren Gründen – nicht auf der Höhe der Zeit, so dass neben der graphischen Benutzerschnittstelle die Konzeption und Implementierung einer neuartigen Satzengine zentrales Ziel des Projekts „XML-Print“ war. Durch den dazu gewählten Ansatz einer funktionalen Programmiersprache kann die Mehrprozessoren-Architektur heutiger Computer durch Parallelisierung sehr gut ausgenutzt werden. Zudem werden Seiteneffekte aufgrund des Aufbaus funktionaler Sprachen verhindert.

Auf Basis des Frameworks Mono⁹ wurde in F#¹⁰ eine plattformunabhängige Software entwickelt. Diese lässt sich nicht nur aus der graphischen Oberfläche von XML-Print heraus aufrufen, sondern auch eigenständig über die Kommandozeile / Shell. Somit sind auch ein Batchbetrieb oder die serverseitige Nutzung ohne weiteres möglich.¹¹

3. Inhalte des Workshops

Der Workshop zeigt Anwendern anhand einer Beispieldition die Arbeitsweise von XML-Print auf. Dazu wird ausgehend von einem XML-Dokument zunächst der grundsätzliche Mechanismus von *Formaten* und *Zuweisungen* erklärt und auf verschiedene Elemente des Quelldokuments angewendet. Sodann werden verschiedene Elemente einer Publikation wie Papiergröße, Seitenränder, Schriftart, Kopf- und Fußzeile oder Spaltenzahl angepasst. Abschließend werden spezielle Strukturen wie Fuß- und Endnoten sowie Apparate integriert.

Den Teilnehmern wird die gesamte Bandbreite von XML-Print veranschaulicht, so dass sie in die Lage versetzt werden, das Werkzeug direkt auch für ihre konkreten Publikationsprojekte einzusetzen. Neben der finalen Druckausgabe können dabei z. B. auch verschiedene Lesefassungen

⁹ Siehe <http://mono-project.com/>.

¹⁰ Siehe <http://fsharp.org/>.

¹¹ Als Beispiel für den Einsatz auf einem Server sei auf die Druckfunktion innerhalb des Deutschen Wörterbuchs des Wörterbuchnetzes verwiesen: <http://woerterbuchnetz.de/DWB/>.

eines Quelldokuments mit variierendem Informationsgehalt für unterschiedliche Zielgruppen erzeugt werden.

4. Teilnehmerkreis / Technische Ausstattung

Der Workshop richtet sich an alle interessierten Wissenschaftler, die ihre XML-Dokumente in ein druckfertiges Format überführen wollen oder auch verschiedene Lese- und Zwischenversionen benötigen. Besondere Vorkenntnisse sind nicht nötig. Da es sich um eine „Hands-on“-Sitzung handelt, sollte die Teilnehmerzahl 25 nicht übersteigen. Die Teilnehmer benötigen einen Laptop, auf dem das Programm vorab installiert wurde.¹² Für die Präsentation wird ein Beamer benötigt.

Literatur

- Burnard, Lou und Syd Bauman (2007). *TEI P5. Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative.
- Extensible Stylesheet Language (XSL)* (2006). *W3C Recommendation 05 December 2006*. Version 1.1.
<http://www.w3.org/TR/xsl11/>.
- Extensible Stylesheet Language (XSL)* (2012). *W3C Working Draft 17 January 2012*. Version 2.0.
<http://www.w3.org/TR/xslfo20/>.
- TextGrid: Digital edieren – forschen – archivieren* (2013). <http://textgrid.de/>.
- XML Path Language (XPath)* (2010). *W3C Recommendation 14 December 2010*. Version 2.0. <http://www.w3.org/TR>xpath20/>.
- XML-Print: typesetting arbitrary XML documents in high quality* (2013). <https://sourceforge.net/projects/xml-print/files/>.

¹² XML-Print ist erhältlich unter: <https://sourceforge.net/projects/xml-print/files/>.

GenericViewer - Semantische Annotation und 3D-Informationen in den Spatial Humanities

Einreichung Dhd Passau 2014, Projekt IBR

Einleitung

Forschungen im Bereich der *Spatial Humanities* arbeiten bisher zumeist in großräumigen geographischen Dimensionen. Dies gilt in besonderer Weise für Vorhaben, die Geographische Informationssysteme (GIS) zur Analyse und Visualisierung von Daten zur Verbreitung von Phänomenen, Entwicklungen und Objekten aus unterschiedlichen Forschungsfeldern wie etwa der Archäologie, der Wirtschaftsgeschichte oder der Raumsoziologie nutzen. Die bearbeiteten Räume sind auf Flächen, nämlich auf geographische Karten, projiziert, die Datenanalyse vorwiegend quantitativ [s. z. B. v. Lünen, Travis 2013].

Dem gegenüber stehen Disziplinen wie etwa die Kunstgeschichte und die Liturgiewissenschaft, deren Untersuchungsgegenstände sich in kleineren Räumen befinden oder selbst kleinere Räume sind, wie Kirchen, Paläste und Häuser, Plätze, Gärten und urbane Strukturen [Ananieva et al. 2013]. Hier ist es notwendig, den dreidimensionalen Raum zu betrachten. Die dritte Dimension spielt in solchen kleineren Räumen naturgemäß eine größere Rolle, die Projektion auf die Fläche führt dementsprechend zu einem größeren Informationsverlust. Allerdings gibt es in den Digital Humanities bisher nur wenige entsprechende Forschungsansätze [cf. Paliou, Knight 2010].

Das Mainzer Projekt “Inscriften im Bezugssystem des Raumes IBR” erarbeitet Wege, 3D-Informationen in den Mittelpunkt raumbezogener geisteswissenschaftlicher Forschung zu stellen. Eine im Projekt entwickelte Erfassungs- und Visualisierungssoftware bietet dabei die Möglichkeit, visuelle und textuelle Informationen zu dreidimensional vermessenen Objekten in ihrem historischen räumlichen Bezugssystem zu erfassen. Die Dichte dieser multimedialen Daten wird mit dem Werkzeug nach Kategorien und Zusammenhängen abfragbar und visuell erfahrbar. Forschungsaussagen gewinnen dadurch an empirischer Qualität.

Software

Panorama-Fotografie ist eine inzwischen weit verbreitete Technik, bei der man Bilder mit sehr großen Blickwinkeln (bis zu 360°) aus einigen Schnappschüssen zusammensetzt. Nicht nur kommerzielle Software wie Google Street View kann solche

Panoramen visualisieren, auch populärwissenschaftliche Anwendungen¹ vermitteln dem Benutzer ein realistisches Raumerlebnis.

Um jedoch raumbezogene Forschung betreiben zu können, ist es zunächst notwendig, den Untersuchungsgegenstand zu vermessen, da einzelne Bilder keine automatisch ableitbaren 3D-Informationen beinhalten. Ein Weg ist die Vermessung mittels terrestrischem Laserscanning (TLS). Damit lassen sich in wenigen Minuten sog. Punktwolken mit hoher Qualität innerhalb eines Bereiches bis zu hundert Metern Entfernung vom Messstandpunkt generieren. Die meisten derzeitigen Programme zur Weiterverarbeitung solcher Punktwolken verlangen jedoch ein sehr großes technisches Hintergrundwissen.

Das vom IBR-Team entwickelte Erfassungs- und Visualisierungswerkzeug „GenericViewer“ bietet typische Funktionen eines Panorama-Viewers, und ermöglicht darüber hinaus, im Panorama-Bild 3D-Objekte zu identifizieren und zu annotieren. Die Software steht damit auch Wissenschaftlern ohne Erfahrung mit komplexeren 3D-Anwendungen zur Verfügung. Dies gilt umso mehr, als es sich um eine leichtgewichtige Web-Anwendung handelt, die als Systemvoraussetzung lediglich einen Browser mit aktiviertem Javascript und WebGL-Unterstützung erfordert.

Datenmodell und Verknüpfung mit textuellen Informationen

Die Identifizierung eines Objektes im GenericViewer beginnt mit der Markierung ihres Umrisses als „Geometrie“ in der Punktwolke. Geometrien erscheinen im Datenmodell zunächst jedoch nur als geordnete Listen von Koordinaten. Um sie für die geisteswissenschaftliche Analyse nutzbar zu machen, müssen sie inhaltlich gekennzeichnet werden. Im GenericViewer geschieht dies durch semantische Annotation mit textuellen und strukturierten Informationen. Die drei Hauptfunktionen von Annotationen sind in diesem System (1) Geometrien als Exemplare einer bestimmten Kategorie, zum Beispiel „Altar“, zu identifizieren, (2) über diesen klassifizierten Geometrien weiter gehende Aussagen in Form von strukturierten (also maschinenlesbaren) Daten zu treffen und (3) diese Daten mit zusätzlichen Informationsressourcen wie anderen Geometrien oder textuellen Quellen zu verknüpfen. Eine Grabinschrift beispielsweise sollte als solche gekennzeichnet und mit einem epigraphischen Fachartikel, der sie beschreibt, verbunden werden können. Annotationen sollten dabei verschiedene Ressourcen, anders als nur rein technisch durch einen Hyperlink, *semantisch* zueinander in Bezug setzen. So sollte etwa die Verknüpfung einer Geometrie mit einer auf sie Bezug nehmenden Textstelle in einem wissenschaftlichen Artikel ihrerseits klassifiziert werden können, etwa als Beitrag zur Datierung oder als quellenkritischer Kommentar. Dies ist zum Beispiel wichtig, um in

¹ Zum Beispiel eine Viewer-Anwendung zum Speyerer Dom: <http://www.kaiserdom-virtuell.de>

einer Suchfunktion effizient alle für eine markierte Inschrift relevanten Textstellen für eine gegebene Fragestellung zu finden.

Aus diesen Gründen ist der GenericViewer als semantische Annotationsumgebung konzipiert worden. Das Datenmodell besteht aus miteinander verknüpften Aussagen der Form Subjekt-Prädikat-Objekt, die als strukturierte Daten im RDF-Format repräsentiert werden. Die Begrifflichkeiten für diese Aussagen können sowohl selbst erstellten projektspezifischen als auch externen Taxonomien und Ontologien wie dbpedia² oder dem Getty Arts and Architecture Thesaurus³ entnommen werden. Texte und Textstellen wiederum lassen sich mit einer angepassten Version des Textannotators Pundit [Grassi et al. 2012] semantisch untereinander oder mit Geometrien verknüpfen. Das entstehende Datenmodell ist eine Repräsentation des Untersuchungsgegenstandes und der sich darauf beziehenden Forschungsdiskurse. Aufgrund der Verwendung von RDF-Repräsentationen sind die Daten Teil des Semantic Web mit automatisiertem Zugriff. Das ermöglicht die Aggregation in größere Datenbestände und die Entwicklung darauf aufbauender neuer Anwendungen.

Fallstudie

Der geschilderte technisch-methodische Ansatz von IBR wird zur Zeit im Rahmen einer Fallstudie zur Liebfrauenkirche in Oberwesel erprobt. Dieser spätgotische Sakralbau ist durch einen in Teilen intakten historischen Innenraum mit einer erhaltenen liturgischen Ausstattung und durch eine gute Dokumentenlage gekennzeichnet. Insbesondere sind die vorhandenen historischen Inschriften im epigraphischen Fachkatalog „Die Deutschen Inschriften“ bzw. in dessen digitalem Pendant, „Deutsche Inschriften Online“ (DIO⁴), kritisch editiert. Von besonderer Bedeutung für die Studie ist das Vorhandensein eines in seinen mittelalterlichen Strukturen aussagekräftigen Raumgefüges. So ergeben sich zum Beispiel durch den erhaltenen Lettner, der den Innenraum in messbare Teilräume untergliedert, ideale Untersuchungsmöglichkeiten hinsichtlich Sichtbarkeits- und Zugänglichkeitsmustern nach dem Spatial-Syntax-Ansatz [Hillier 1999, Clark 2010].

Zu den in der Fallstudie behandelten Ausstattungsstücken gehören Grabmonumente, die im Zusammenhang von Kapellen und Altarstiftungen dargestellt werden. Dabei wird eine Zuweisung bestimmter Bereiche der Kirche zu politischen und gesellschaftlichen Gruppen vorgenommen, außerdem werden Funktionsbereiche für Liturgie, Andacht und gesellschaftliche Handlungen markiert und mit relevanten Text- und Bildquellen verbunden. Unter anderem wird die Sichtbarkeit bestimmter Inschriften und Altäre für verschiedene Standpunkte mit Hilfe des

² <http://de.dbpedia.org/>

³ <http://www.getty.edu/research/tools/vocabularies/aat/index.html>

⁴ www.inschriften.net

Erfassungswerkzeugs untersucht, um so genauere Aussagen über die Bedeutung von Teilräumen, Taburäumen und über Sichtbarrieren wie den Lettner zu treffen. Beispielsweise gibt es in der Inschriftenforschung die These, dass die Gründungsinschrift von Liebfrauen nur vom Standpunkt vor dem Hauptaltar vollständig erkennbar war [cf. Nikitsch 1996]. Dies wäre von Bedeutung - so die These - da eine im Text vorhandene politische Anspielung sich an den die Kirche am Hauptaltar weihenden Bischof richtete. Der GenericViewer bietet hier die Möglichkeit, sich dem Untersuchungsgegenstand intuitiv zu nähern, die geschilderte Vermutung durch eine Sichtbarkeitsanalyse empirisch zu prüfen und das Ergebnis am dreidimensionalen Objekt in Form strukturierter Daten zu dokumentieren.

Diskussion

IBR stellt raumbezogene Forschungsfragen im Kontext epigraphischer Untersuchungen. Gleichwohl trägt die im Projekt entwickelte Software bewusst den Anspruch der Generizität im Namen. Denn es können mit dem GenericViewer Fach- und Messdaten beliebigen Inhalts dargestellt und miteinander verknüpft werden. Dies eröffnet vielfältige weitere Nutzungsmöglichkeiten auch über die geisteswissenschaftliche Forschung im engeren Sinne hinaus. Zu denken wäre etwa auch an denkmalpflegerische Anwendungen sowie an die Katalogisierung von Artefakten und die Verknüpfung mit Europeana.

Zum Zeitpunkt der Niederschrift ist die Verknüpfung von Panoramafotos nur mit Punkt wolken aus Laserscans möglich. Die Genauigkeit eines solchen Scans ist für geisteswissenschaftliche Fragestellungen häufig nicht erforderlich, daher sollten kostengünstigere Techniken wie *Structure from Motion*⁵ in der weiteren Entwicklung unterstützt werden. Gleches gilt für einfache Analysen: Trotz standardisierter Schnittstellen sind zur Verwendung externer Software häufig kleine Anpassungen oder Formatierungen notwendig. Daher sollten Sichtbarkeitsanalysen und Abfragen auf dem Datenbestand stärker in den GenericViewer integriert werden, so dass der Einsatz spezieller Programme nur noch in komplexeren Szenarien nötig ist. Insbesondere Sichtbarkeit ist von einer Vielzahl komplexer Faktoren abhängig, die nicht alle durch den GenericViewer berücksichtigt werden können. Einige Sichtbarkeitsfragen, z. B. nach der Möglichkeit, menschliche Handlungen aus der Distanz wahrnehmen zu können, können sogar nur mit menschlicher Unterstützung beantwortet werden [Clark 2012, S. 87 ff.].

IBR legt Wert auf eine hohe Interoperabilität von Ressourcen, die durch Veröffentlichung von Quellcode (*Open Source*) und die Einhaltung von Standards erreicht wird. Zusammenarbeit, Erweiterung des Quellcodes, die Anbindung von externen Anwendungen, Analysewerkzeugen oder Ontologien werden gewünscht und unterstützt. Dessen ungeachtet hängt die Nützlichkeit und die Nachhaltigkeit der mit

⁵ s. http://en.wikipedia.org/wiki/Structure_from_motion

dem Annotationswerkzeug produzierten Daten entscheidend von der Wahl der Ontologien ab. Nutzer stehen hier stets vor einem Ausgleich zwischen Spezifität und Verallgemeinerbarkeit. Außerdem spielt die Form des zu Grunde liegenden digitalen Textdokumentes eine entscheidende Rolle. Der Textannotator Pundit ist wie ähnliche Tools [Khalili et al. 2012] auch darauf ausgelegt, Positionen in HTML-Seiten zu annotieren. Diese stellen jedoch eher die Präsentationsschicht eines Webdokumentes als den eigentlichen Text mit seinen Gliederungselementen dar. Aus diesem Grund bietet der GenericViewer Sonderfunktionen für TEI-XML-Dokumente. Eine Klärung der Frage nach dem geeigneten Format für austauschbare und nachhaltige Annotationen bleibt aber ein Desiderat der Digitalen Geisteswissenschaften insgesamt.

Im Kontext der Konferenz ist schließlich auf die Frage nach dem analytischen Mehrwert des von IBR entwickelten digitalen Werkzeugs zu antworten: Semantische Annotation ist eine Kommunikationspraxis, keine Analysemethode. Auch durch die Verarbeitung und Visualisierung von 3D-Daten allein werden noch keine neuen Forschungsergebnisse erzielt. Zwar erlangen analytische Begriffe wie „Zentralität“ oder „Verbundenheit“ durch vom GenericViewer unterstützte Metriken, wie sie etwa die *Visibility Graph Analysis* [Turner et al. 2001] bereitstellt, ein breiteres empirisches Fundament. Diese Ergebnisse, sowie andere, nicht metrifizierbare raumbezogene Aussagen lassen sich jedoch erst durch eine geeignete visuelle Darstellung wirklich verständlich machen. Das zeigt aber auch, dass die Grenze zwischen Visualisierung und Analyse durch technische Systeme insbesondere in der nichtquantitativen geisteswissenschaftlichen Forschung fließend sein kann.

Literatur

- Ananieva, Anna; Bauer, Alexander; Leis, Daniel; Morlang-Schardon, Bettina; Steyer, Kristina (Hg.): Räume der Macht. Metamorphosen von Stadt und Garten im Europa der Frühen Neuzeit. Bielefeld: transcript Verlag, 2013
- Clark, David L. Chatford (2007): Viewing the Liturgy: A Space Syntax Study of Changing Visibility and Accessibility in the Development of the Byzantine Church in Jordan. In: World Archeology, Vol. 39 No. 1 (Mar. 2007), S. 84-104.
- Grassi, Marco; Morbidoni, Christian; Nucci, Michele; Fonda, Simone; Ledda, Giovanni (2012): “Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries”. In: Mitschick, Annett; Loizides, Fernando; Predoiu, Livia; Nürnberger, Andreas; Ross, Seamus (Hg.): Semantic Digital Archives 2012. Proceedings of the Second International Workshop on Semantic Digital Archives (SDA 2012), Paphos, Cyprus, September 27, 2012, CEUR-WS.org/Vol-912.
- Hillier, Bill: Space is the Machine: A Configurational Theory of Architecture, Cambridge: University Press, 1999.

- Khalili, Ali ; Auer, Sören ; Hladky, Daniel (2012): The RDFA Content Editor - From WYSIWYG to WYSIWYM. In: Proceedings of COMPSAC 2012 - Trustworthy Software Systems for the Digital Society, July 16-20, 2012, Izmir, Turkey, 2012.
- Nikitsch, Eberhard Josef (1996): Ein Kirchenbau zwischen Bischof und Stadtgemeinde. Zur angeblich verlorenen Bauinschrift von 1308 in der Liebfrauenkirche zu Oberwesel am Rhein, in: JbWdtLg 22 (1996). S. 95-112.
- Paliou, E. und Knight, D.J. (2010): Mapping the Senses: Perceptual and Social Aspects of Late Antique Liturgy in San Vitale, Ravenna. In: Contreras, F.; Farjas , M. und Melero , F.J. (Hg.): Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, CAA2010.
- Turner; Doxa, M., O'Sullivan, D., and Penn, A. (2001): "From isovists to visibility graphs: a methodology for the analysis of architectural space". Environment and Planning B 28 (1). S. 103-121.
- v. Lünen, Alexander und Travis, Charles: History and GIS. Epistemologies, Considerations and Reflections. Heidelberg [u.a.]: Springer, 2013.

Projektpräsentation ePol – Postdemokratie und Neoliberalismus

Matthias Lemke / Gregor Wiedemann

Die Postdemokratiediagnose, wonach sich repräsentative Demokratien gegenwärtig unter dem Druck einer neoliberalen Hegemonie zunehmend vom Ideal politischer Teilhabe entfernen, kann auf unterschiedlichen Wegen empirisch überprüft werden. Das ePol-Projekt zielt darauf ab, anhand des Nachweises einer zunehmenden Ökonomisierung von Sprache in der politischen Öffentlichkeit einen Beleg für aktuelle Tendenzen demokratischer Degeneration zu liefern. Ausgangspunkt hierfür ist ein Corpus von 3,5 Millionen Zeitungsartikeln deutschsprachiger Qualitätszeitungen aus dem Zeitraum von 1947 bis 2012, der pars pro toto für die bundesrepublikanische politische Öffentlichkeit steht. In ihm suchen wir nach Dokumenten und Aussagen bis hinunter zur Satzebene, die im Zuge der Plausibilisierung politischer Entscheidungen maßgeblich auf marktaffine Inhalte abstellen. Wörterbuchbasierte Retrievalverfahren zur Dokumentidentifikation werden dabei durch manuelle Annotationsverfahren ergänzt. Zudem erlauben Topic-Modelle sowie Kookkurrenzberechnungen eine Einschätzung darüber, inwieweit im Zeitverlauf tatsächlich Konjunkturen einer Ökonomisierung des Politischen – und damit einer Postdemokratisierung repräsentativer Demokratien – festzustellen sind.

Im Rahmen unserer Präsentation erläutern wir grundsätzliche politiktheoretische sowie methodologische Aspekte der Suchstrategie und stellen erste Ergebnisse vor.

Presentation of ePol – Postdemocracy and Neoliberalism Project

Matthias Lemke / Gregor Wiedemann

The diagnosis of post-democracy, which holds that today's representative democracies loose increasingly touch to the ideal of political participation due to the pressure of neo-liberal hegemony, can be tested empirically in different ways. The ePol project aims to provide a proof of current tendencies of democratic degeneration based on the detection of an increasing market orientation in everyday language as it is used in the public sphere. Therefore, we started analyzing a corpus of 3.5 million newspaper articles of German quality newspapers from 1947-2012, which can be considered as an archive representing the public discourse of (West) German political public. The search in the corpus can identify documents and statements down to the level of single sentences and ngrams. It focusses mainly on political decisions described with the help of market-logic or economy-driven speech. Dictionary-based retrieval processes for document identification are combined with manual annotation procedures. In addition, topic models and calculation of word-cooccurrences allow an assessment of the extent to which over time actually conjunctures an economization of politics can be observed. This might indicate in how far a post-democratization of representative democracies has already taken place.

As part of our presentation, we explain basic assumptions from a political theory perspective as well as methodological aspects of our search. Additionally, we provide initial results of our work.

DHd 2014, 25.-28. März 2014, Passau

Abstract zur POSTERPRÄSENTATION (online):

Das Projekt „VerbaAlpina“

Thomas Krefeld, Institut für Romanische Philologie, LMU München

Stephan Lücke, IT-Gruppe Geisteswissenschaften, LMU München

Der Alpenraum stellt eine hinsichtlich natürlicher Gegebenheiten und Lebensbedingungen homogene Zone im Zentrum Europas dar. Von Nizza am Mittelmeer bis Wien an der Donau finden sich die gleichen Geländeformationen wie Täler, Gipfel, Pässe, Kare etc. sowie eine weitgehend einheitliche gebirgsspezifische Flora und Fauna. Neben dieser naturräumlichen Homogenität - und durch diese bedingt - tritt eine kulturräumliche: Die Menschen waren hier traditionell mit ähnlichen Rahmenbedingungen und nicht selten Herausforderungen konfrontiert, denen sie vielfach in ähnlicher, nicht notwendig aber identischer Weise begegnet sind. Als Beispiel sei hier nur Almwirtschaft erwähnt, die im ganzen Alpenraum in vergleichbarer Weise betrieben wurde und wird.

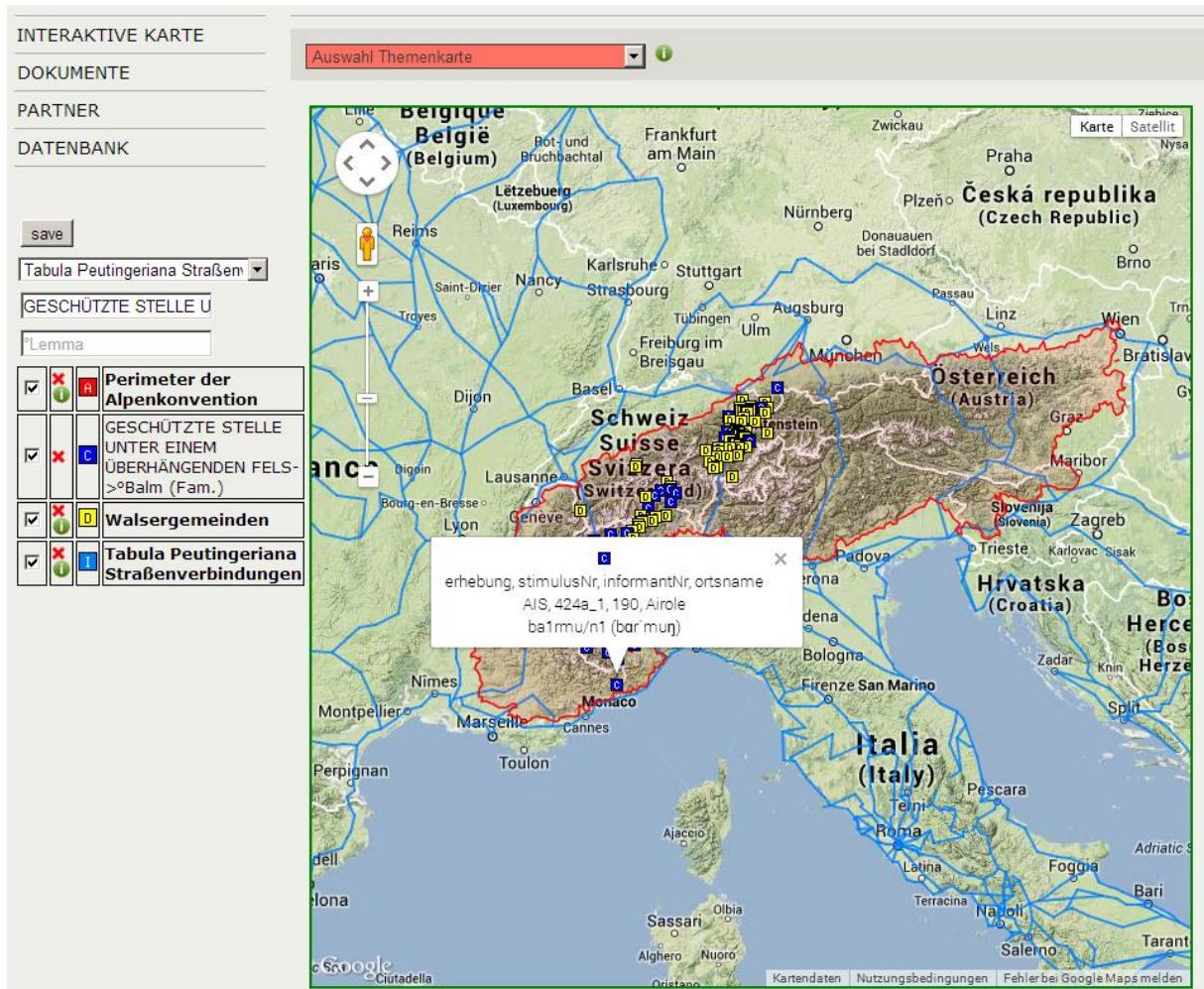
Bei aller Homogenität im beschriebenen Sinn zeichnet sich der Alpenraum gleichzeitig durch eine vielfältige Fragmentierung aus. In ihrer Gesamtheit selbst Grenze zwischen Mittel- und Südeuropa sind die Alpen in ihrem Inneren durch eine Vielzahl von Grenzen unterschiedlicher Art förmlich zersplittet. Dies beginnt bei den einzelnen Talschaften, die aufgrund nur schwer zu überwindender Gebirgszüge wenigstens in früheren Jahrhunderten weitgehend von einander isoliert waren und endet bei den Grenzen der modernen Nationalstaaten, die diesen homogenen Naturraum durchschneiden.

Neben diesen und einer Vielzahl weiterer Grenzen unterschiedlicher Art durchziehen Sprachgrenzen den Alpenraum. Hier treffen Sprachfamilien - das Romanische, das Germanische und das Slavische -, Einzelsprachen (Deutsch, Italienisch, Französisch, Slovenisch) und Dialekte (z.B. Walserdeutsch, Tirolerisch, Gadertalisch, Lombardisch etc.) aufeinander.

Das Projekt VerbaAlpina, das seit gut einem Jahr in einer Kooperation zwischen dem Institut für Romanische Philologie und der IT-Gruppe Geisteswissenschaften der LMU vorangetrieben wird, widmet sich speziell diesem Aspekt alpiner Diversität und konfrontiert ihn mit der ontologischen Homogenität, die aus den natürlichen Gegebenheiten resultiert.

Dabei ist VerbaAlpina in doppelter Hinsicht innovativ: Zum einen durchbricht es aus fachwissenschaftlicher Perspektive die sprachwissenschaftlich häufig isolierte Betrachtungsweise, die sich traditionell aus der überwiegend an den Nationalgrenzen orientierten Sprachdokumentation in Form von Sprachatlanten und Wörterbüchern ergab. Zum anderen setzt es methodisch konsequent auf den Einsatz von DH-Konzepten, was, soweit wir sehen, zumindest für den Bereich der Sprachwissenschaft in der von uns gewählten Form bislang einzigartig ist.

VerbaAlpina ist konzeptionell und technisch bereits sehr weit entwickelt und deutlich über den Status eines bloßen Vorhabens hinaus. Auf dem Projektportal unter <http://www.verba-alpina.gwi.uni-muenchen.de> sind bereits erste Ergebnisse in Form von interaktiven Onlinekarten einsehbar:



Technologisch basiert VerbaAlpina auf dem bekannten Client-Server-Prinzip und setzt überwiegend auf Webtechnologien; die Entwicklung proprietärer Module wie etwa zu installierende Softwareclients wird konsequent vermieden. Sämtliche Daten werden zentral in einer relationalen Datenbank (MySQL) gehalten und gepflegt. Eine PHP-Schnittstelle sorgt für die Präsentation der Daten im Internet. Zentrales Präsentationsmedium der Analysedaten sind die bereits erwähnten Onlinekarten. Deren Erzeugung und interaktive Funktionalität erfolgt aktuell unter Verwendung von Google Maps und durch den Einsatz von Javascript, wobei zu betonen ist, dass das Projekt nicht von Googles Kartendienst abhängig ist. Als Alternative ist die Verwendung von Openstreetmap-Karten möglich.

Inhaltlich ist das Projekt in drei große, voneinander getrennte thematische Bereiche gegliedert, die in zeitlich aufeinander folgenden Phasen abgearbeitet werden:

- Phase 1: Traditionelle Lebenswelt: Almwesen, volkstümliche Medizin, traditionelle Küche
- Phase 2: Natur: Landschaftsformation, Wetter, Fauna, Flora
- Phase 3: Moderne Lebenswelt: Ökologie, Tourismus

Das Basiskonzept besteht dabei aus der Gegenüberstellung von Konzepten (= Begriffen) und Bezeichnungen sowie der georeferenzierten Verteilung dieser Daten im Raum. Die Datenbasis stammt zunächst aus den einschlägigen Sprachatlanten, die für einen Großteil des Alpenraumes verfügbar sind, sowie aus ortsspezifischen Wörterbüchern. Die Datenerfassung muss überwiegend manuell erfolgen, da zum einen die Eigenheiten der speziellen Kartendarstellungen den Einsatz von

OCR verbieten und zum anderen notwendige Kategorisierungen (z.B. Unterscheidung zwischen lexikalischem Lemma und abstrahiertem Worttyp) nur von Spezialisten vorgenommen werden können. In einem zweiten Schritt sollen eventuelle Datenlücken und -inkonsistenzen dann durch gezielte Nacherhebungen ausgeglichen werden. Zu diesem Zweck ist auch an den Einsatz von sog. Social Software gedacht. Zusätzlich zum Sprachmaterial werden in der Datenbank auch außersprachliche Daten zu Geschichte, Ethnographie und Infrastruktur gespeichert, da sich diese Größen einerseits und sprachliche Phänomene andererseits wechselseitig bedingen. Abgerundet wird die Datensammlung durch eine, teils georeferenzierte, Fotodokumentation zur Illustration der im Korpus gesammelten Begrifflichkeiten.

Die Kartenoberfläche gestattet dem Nutzer die interaktive Auswahl und freie Kombination von Konzepten, Bezeichnungen und außersprachlichen Daten im Sinne heuristischer Informationsgewinnung.

Für VerbaAlpina wurde von den Autoren ein Antrag auf Förderung bei der DFG eingereicht, konzipiert als Langzeitprojekt für eine Laufzeit von drei mal drei Jahren. Der Antrag befindet sich aktuell in der Begutachtungsphase. Mit einer Entscheidung wird für Anfang 2014 gerechnet. Unabhängig vom Erfolg dieses Antrags ist beabsichtigt, das Projekt im Rahmen der gegebenen Möglichkeiten weiter zu verfolgen.

Objekte – Raum – Zeit: Die Archäologie – ein Sonderfall der Digital Humanities?

Im deutschsprachigen Raum gehören die unterschiedlichen archäologischen Disziplinen in der Regel zum Fächerkanon der Geisteswissenschaften. Im Rahmen der neu entstehenden Studiengänge oder Center für Digital Humanities wird die Archäologie jedoch häufig ausgespart und es findet eine Konzentration auf die textwissenschaftlich arbeitenden Fächer statt.

Auch im Bereich der Kongresse und Workshops lässt sich diese Entwicklung beobachten, so existiert seit Jahren neben den etablierten Veranstaltungen der Digital Humanities die CAA (Computer Applications and Quantitative Methods in Archaeology), die ausschließlich das Feld der Archäoinformatik abdeckt. Schnittmengen zwischen diesen beiden Gruppierungen entstehen selten.

Im Folgenden soll den Gründen für diese Entwicklung nachgegangen werden.

Eine der Antworten auf diese Frage stellt die Vermutung dar, dass es sich bei der Archäologie überhaupt nicht um eine Geisteswissenschaft handelt. Gestützt wird diese Vermutung dadurch, dass im angelsächsischen Raum die Archäologie meist nicht den Geisteswissenschaften zugeordnet wird. Tatsächlich sind die herangezogenen Quellen, Methoden und Vorgehensweisen gänzlich andere als bei den übrigen Geisteswissenschaften. So stehen nicht Texte und die darin enthaltenen Informationen im Zentrum der Forschung, sondern Objekte und deren Verbindung mit Raum und Zeit bilden das Rückgrat des Erkenntnisgewinns. Die herangezogenen Methoden sind vielfältig und stammen aus einer Vielzahl von Wissenschaften von der Geographie, über die Physik, bis hin zu Mathematik und Statistik. Eine interdisziplinäre Ausrichtung der Vorhaben ist somit meist generisch, eine Beherrschung sämtlicher Methoden durch die Archäologie ist heute nicht mehr zu realisieren.

Man könnte somit zurecht behaupten, dass es sich bei der Archäologie nicht um eine Geisteswissenschaft handelt, sondern vielmehr um eine Verbindung aus Natur- und Geowissenschaften. Dieser Ansatz greift jedoch zu kurz, er lässt außer Acht, dass sich diese Beobachtung lediglich auf die Methoden zur Datengenerierung beziehen. Eine Ausgrabung mit all ihren Anforderungen kann natürlich tatsächlich ausschließlich durch technisches Personal durchgeführt werden. Hieraus entstehen jedoch lediglich Pläne, Tabellen, Fotografien, Datenbanken sowie eine Unzahl von Artefakten und Befunden. Einen Einblick in die Kulturgeschichte des untersuchten Ortes und Zeitraumes ergibt sich hieraus nun nicht von selbst. An dieser Stelle ist nun der geisteswissenschaftliche Sachverstand gefordert, der die Ergebnisse ordnen, sortieren und mit anderen Quellen in Einklang bringen kann, um auf diese Weise neues Wissen aus den Daten zu generieren.

Die Archäologie ist somit in ihrer Methodik häufig den anderen Geisteswissenschaften fremd, in ihren Ergebnissen und Schlüssen dann jedoch wieder ganz eine von ihnen. Ein Ausschluss dieser Wissenschaft kann somit nicht durch ihre generische Andersartigkeit, sondern vielmehr durch gegenseitiges Unverständnis und unterschiedlich gewachsene Strukturen erklärt werden.

Wenn die Archäologie nun einen Platz im Rahmen der Digital Humanities beansprucht, ist die Frage berechtigt, was kann sie hierzu beitragen und wo liegen die Anknüpfungspunkte?

Im Gegensatz zu den übrigen Geisteswissenschaften ist die Archäologie, im besonderen Maße die Feldarchäologie, mittlerweile ohne digitale Unterstützung nicht mehr denkbar. Die Verwaltung von teils mehreren Hunderttausend Artefakten ist mit den heutigen zeitlichen und finanziellen Rahmenbedingungen nicht mehr zu realisieren. Von diesem Wissen um die Verwaltung und Analyse von großen Datenbeständen aus unterschiedlichsten Quellen können selbstverständlich

auch andere Fachbereiche profitieren. Im besonderen Maße trifft dies auf die räumlichen Daten zu, deren Möglichkeiten weit über das punktförmige Kartieren von Daten auf Google-Karten hinausgehen und durch einen Großteil der Geisteswissenschaften noch vollkommen unerkannt sind, während die Analyse und die Visualisierung komplexer räumlicher Zusammenhänge mittlerweile zu einer der Standardmethoden der Archäologie gehört. Immer leistungsfähigere Geoinformationssysteme sowie immer präzisere Daten zu Landschaftsmorphologie, Vegetation, Hydrologie oder Geologie lassen gänzlich neue Analysen zu und führen zu erstaunlichen Ergebnissen. Hochauflösende dreidimensionale Modelle der Erdoberfläche sollen dies anhand einiger Beispiele verdeutlichen.

Die fachlichen Anknüpfungspunkte sind mit den antiken Texten seit jeher gegeben, sie haben nur in den vergangenen Jahren zugunsten neuer Fragestellungen an Bedeutung verloren. Die Entdeckung von Mykene und Troja wären ohne die Analyse der homerischen Epen wohl kaum geglückt, ohne Pausanias Beschreibung von Olympia würden wir wohl nicht wissen, welche Bauten welchem Zweck dienten.

Es wäre zu hoffen, dass wieder mehr Vorhaben diese Schnittstelle besetzen, um ein reiches und differenzierteres Bild der Vergangenheit zu zeichnen, was nur durch eine Verbindung sämtlicher Disziplinen gelingen kann. Virtuelle Arbeitsumgebungen und standardisierte Dateiformate bieten uns hier bisher ungeahnte Möglichkeiten, Text, Objekt, Raum und Zeit miteinander zu verbinden und erstmals gemeinsam zu visualisieren und zu analysieren. Es ist zu hoffen, dass sich Vertreter der einzelnen Disziplinen aufeinander zubewegen, um gemeinsam neuen Erkenntnisse mit neuen Methoden zu generieren.

Die Archäologie stellt zwar durch ihre Methodik einen Sonderfall innerhalb der Geisteswissenschaften dar, gehört jedoch unzweifelhaft zu diesen und muss daher auch innerhalb der Digital Humanities einen Platz beanspruchen. Gegenseitig können die unterschiedlichen Wissenschaften nur voneinander profitieren, gerade auf die digitalen Methoden und Vorgehensweisen trifft dies im besonderen Maße zu.

Workshop-Vorschlag für die Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd) vom 25.–28.03.2014 an der Universität Passau

Digital Humanities – methodischer Brückenschlag oder „feindliche Übernahme“? Chancen und Risiken der Begegnung zwischen Geisteswissenschaften und Informatik

Gehör verschaffen! Die Produktion und Verbreitung von digitalen Radiosendungen

Podcasts, digitale Radiosendungen oder abbonnierzbare Audioinhalte sind eine Möglichkeit der Verbreitung von wissenschaftlichen Inhalten, die seit einiger Zeit einen Boom erleben. Die Gründe dafür sind vielfältig, lassen sich aber vermutlich vor allem in den sich verändernden Rezeptions- und Produktionsbedingungen von digitalen Inhalten verorten. Das zeigt sich beispielsweise in der mobilen Nutzung durch Smartphones (Rezeption) und in der zunehmenden Leistbarkeit und Anwendbarkeit von technischem Equipment (Produktion).

Podcasting ist dabei wesentlich mehr als nur die Publikation von Audiodateien im Internet. Es ermöglicht Kommunikation von Wissen, häufig in Dialogform vermittelt, und stellt über die Stimme gleichzeitig einen persönlicheren Bezug zum Publikum her, als schriftliche Inhalte. Als Podcasting kann zwar auch das Bereitstellen von Vorträgen auf einer Konferenz gelten, etwa für Personen, die nicht persönlich anwesend sein können oder zur langfristigen Dokumentation der Präsentationen. Das Potential von Podcasts oder digitalen Radiosendungen erschöpft sich darin aber lange nicht. Vielmehr lassen sich mit der Aufzeichnung von Gesprächen und Diskussionen im Umfeld einer Konferenz Themen, Vorträge und Debattenbeiträge verdichten und vertiefen, etwa für ein spezifisches Fachpublikum, das auf diese Weise verstärkt in Dialog treten kann. Mehr noch können – je nach konzeptioneller Rahmung – Podcasts auch einen Einstieg in bestimmte Themenfelder ermöglichen, etwa für ein externes, interessiertes Publikum oder auch beispielsweise für Studierende, die sich in ein Thema einarbeiten wollen. Das heißt, Podcasting ermöglicht Wissensaustausch und Kommunikation auf vielen unterschiedlichen Ebenen, von interner Projektkoordination, über ausführliche, externe Projektdokumentation und -diskussion bis hin zur Ergebnis-Präsentation für ein breites, interessiertes Publikum. Dadurch wird ein Diskussionsraum geschaffen, der neue Formen von Feedbackschleifen ermöglicht.

Der Workshop zur Produktion und Verbreitung von digitalen Radiosendungen richtet sich an Personen, die einen Einstieg in das Thema Podcasting suchen und erfahren möchten, welche Möglichkeiten der Produktion und Distribution es gibt und, die Interesse daran haben, sich an der Aufnahme von Sendungen zu beteiligen. Vorwissen wird nicht vorausgesetzt. Prinzipiell ist die Teilnahme auch ohne eigenes technisches Equipment möglich, es wäre allerdings von Vorteil, wenn vorhandene Field-Recorder, Mikrofone, Laptops, Tablets oder Smartphones mitgebracht werden würden.

Das Programm des Podcasting-Workshops gliedert sich in drei Teile, wobei nur die ersten beiden Teile offizielle Programmpunkte darstellen. Der dritte Teil soll ein Podcasting-Testlauf unter Realbedingungen auf der Digital Humanities-Jahrestagung werden. Das heißt, in der ersten Hälfte des Workshops sollen die theoretischen Grundlagen geschaffen werden für den zweiten Teil des Workshops, wo es um die Planung, Produktion und Veröffentlichung einer Podcastepisode geht. Es wäre wünschenswert, wenn im Anschluss an den Workshop einige Teilnehmer/Teilnehmerinnen Interesse an der Herstellung von Podcasts von der DHd-Tagung zeigen würden. Das würde nicht nur der besseren Dokumentation der Konferenz dienen, sondern gleichzeitig das Potential von Podcasting zur Kommunikation von wissenschaftlichen Inhalten in einem Praxistest aufzeigen.

1. Einführung: Was ist ein Podcast?

Im ersten Teil des Workshops wird es um die theoretische Vorbereitung gehen. Neben der Frage, wie Podcasting entstanden ist, wird zu klären sein, was ein Podcast ist, wie Podcasts gehört werden können, z.B. durch das Abonnieren von RSS-Feeds. Anschließend ist geplant, den Weg einer Podcastproduktion von der Planung bis zur Veröffentlichung zu verfolgen. Dabei wird über Sendungsformate diskutiert, Aufnahmetechniken (Software und Hardware) vorgestellt und auf unterschiedliche Distributionsformate und -kanäle eingegangen. Insbesondere soll die Bedeutung von einheitlichen Metadaten für Audio- oder Videodateien aufgezeigt werden.

2. Podcastproduktion: Gehör verschaffen!

Im zweiten Teil des Workshops geht es um die Umsetzung der im 1. Teil erarbeiteten Grundlagen des Podcastings. Alle Teilnehmer/Teilnehmerinnen sollen (einzelne oder in kleineren Gruppen) eine Podcastfolge planen und anschließend aufnehmen und veröffentlichten – Veröffentlichung muss dabei nicht unbedingt heißen, dass das Audiofile frei zugänglich ist. Zur internen Projektkommunikation wäre es genauso denkbar, dass ein Audiofile nur innerhalb einer begrenzten Personengruppe zirkuliert. Feeds lassen sich zum Beispiel mit Passwort schützen und Audiofiles können auch über einen gemeinsamen Ordner auf einem Server verteilt werden.

Sollte der Workshop-Vorschlag für die DHd-Tagung angenommen werden, müsste im Vorfeld noch geklärt werden, in welcher Form eine Veröffentlichung stattfinden könnte und entsprechende Infrastruktur vorbereitet werden. Denkbar wäre unter anderem eine Veröffentlichung auf dem Blog zur DHd-Jahrestagung oder auch auf einer eigens für den Podcasting-Workshop eingerichteten Website. An der Stelle wäre zunächst nicht viel mehr nötig, als ein Blog einzurichten, schließlich ist aus technischer Perspektive ein Podcast lediglich ein Feed inklusive *Media Enclosure*. Die Dokumentation des Workshops könnte dann in Zukunft eine erste Anlaufstelle sein für Personen im Umfeld der Digital Humanities, die sich für die Produktion von Podcasts interessieren und Audioaufnahmen planen.

3. Podcasting während der Konferenz

Der dritte Teil des Workshops erstreckt sich über die gesamte Zeit der Konferenz (und vielleicht auch darüber hinaus) und ist mit der Hoffnung verbunden, dass sich einige Teilnehmer/Teilnehmerinnen des Workshops an der Produktion von Sendungen über die DHd-Tagung beteiligen. Die Produktion von digitalen Radiosendungen sollen einen Kommunikationsraum eröffnen, in dem Inhalte und Themen der Digital Humanities vorgestellt und diskutiert werden. Durch unterschiedliche Sendungsformate sollen in dem Fall zwei Kommunikationsebenen angesprochen werden: Erstens die Vermittlung von Inhalten für Kollegen und Kolleginnen, die nicht vor Ort sein können und zweitens für interessierte Personen, für die die Podcasts eine Einstiegshilfe in das Themenfeld Digital Humanities bieten sollen. Für den Fall einer Annahme des Workshop-Vorschlags wäre noch zu überlegen, inwiefern es sinnvoll wäre, einige Gespräche bereits vorab zu führen und zu veröffentlichen – beispielsweise als Teaser oder um interessante Vorabinformationen zu bieten.

Für den dritten Teil des Workshops wäre ebenfalls die Unterstützung durch die Veranstalter notwendig. Das betrifft vor allem die Räumlichkeiten, denn es müsste ein ruhiger Raum zur Verfügung gestellt werden, in dem die Podcasts aufgenommen werden könnten. Konkrete Sendungsformate würden im zweiten Teil des Workshops erarbeitet werden, wofür sich folgende Gesprächscluster anbieten würden:

- Interviews mit Vortragenden (z.B. Zusammenfassung des Vortrags, Diskussion)
- Interviews mit Personen aus dem Organisationsteam (z.B. Hintergründe zur Tagungsorganisation und DHd)
- Interviews mit Personen, die im Bereich Digital Humanities arbeiten, aber auf der Jahrestagung keine eigene Präsentation haben
- Diskussionsrunden zu Kern- und Nischenthemen der Digital Humanities

Creating dictionaries for argument identification by reference data

The creation of dictionaries is an important task to conceptualize and operationalize research questions in content analysis (Neuendorf, 2002). One can define concepts for coding operationalized variables in the form of mutual exclusive categories or decide if the content of documents is relevant for coding within the research task by the formalization of meaning through a dictionary (Krippendorff, 2004). Dictionaries are often defined on the basis of a “theory of meaning that reflects a research question or the vocabulary of an academic discipline” (Krippendorff, 2004). Thus, we can think of dictionaries as operationalized representations of historical, sociological, cultural or political theories that are investigated within humanities research.

In contrast to manual dictionary creation from a small set of selected sample documents we present our approach to automatically extract dictionaries from a reference corpus of arbitrary size. For our goal of identifying arguments for a political science research task we create two dictionaries. One semantic dictionary on the utilization of topic models (Blei/Ng/Jordan, 2003; Teh/Jordan, 2010) to identify thematically relevant documents; and one rather syntactical dictionary based on term similarities of linguistic markers to identify a high density of argument structures. We present the idea, results and an example application of the extracted dictionaries for relevancy judging of retrieval results in large digital document collections.

Semantic dictionaries via topic models

Domain experts easily can compile a small reference corpus of paradigmatic documents containing contents of their interest. On this reference corpus we apply a topic model based on the Pitman-Yor Process (Teh, 2006). It employs Poisson instead of Dirichlet distributions which better approximate distributions of natural language data. One of the key properties of topic modeling is the inference of not directly observable variables considered as latent topics. A distribution over these latent topics (classes of co-occurring terms) is allocated to each of the documents within a digital text collection. Another hidden variable describes each of those topics in form of a probability distribution over the vocabulary of the text collection. On the basis of the assumption that all of the topics, extracted in a certain abstraction level controlled by the model parameters, represent the meaning and content of a digital text collection in a compressed form we created our dictionary extraction process. For this process we utilize the set of all resulting topics \mathbf{z} to calculate scores for each word in the vocabulary within the collection. Since we have the property that only a few terms in a topic have high probability we use only a limited number of the most probable words in each topic. The score for each word is calculated by

$$score(w_n) = \log(F(w_n)) \sum_{k=1}^K p(w_n|z_k) ,$$

where $p(w_n|z_k)$ is the probability of the n th word in the vocabulary within the k th topic of the model and $F(w_n)$ is the absolute word frequency of the term w_n within the text collection. The idea behind this formula is that terms of high probability within a topic have significance for the meaning of the text collection. Furthermore we take the frequency of the word into account

because a high frequent use of a term and a high probability within a topic induce a prototypical usage within the texts. Using topic models further allows for filtering of unwanted semantical structures when creating dictionaries from the collections. In our application we identified a foreign language 'topic' and a topic thematically not related to our research question in the reference texts and could easily exclude them from our k topics before applying the score calculation.

Syntactic dictionaries via term similarities

Additionally to our dictionary containing semantic information related to theoretical aspects of the political science research task we created a second dictionary of linguistic markers which can be employed to identify argumentative structures. We took a list of 46 German linguistic markers from another research project on causality and textual coherence (Breindl/Walter, 2009) as a starting point. This list was incrementally extended up to 144 terms by automatically computed synonyms of the markers retrieved from the database of the "*Projekt Deutscher Wortschatz*" (Quasthoff/Eckart, 2009), a representative corpus of German language.

Application

We applied these dictionaries for retrieval of documents in a large collection of newspaper articles to identify argumentative texts with a certain ideological framing. The retrieved texts then are subject of a close reading process of political scientists which utilize the dictionaries for qualitative coding schemes.

References

- Alsumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic Significance Ranking of LDA Generative Models. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I (S. 67–82). Berlin, Heidelberg: Springer-Verlag.
- Breindl, Eva / Walter, Maik (2009): Der Ausdruck von Kausalität im Deutschen. Amades - Arbeitspapiere zur deutschen Sprache, Mannheim.
- Blei, D.M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3, 993–1022.
- Krippendorff, K. (2004). Content analysis: an introduction to its methodology (2nd ed.). Thousand Oaks Calif.: Sage.
- Neuendorf, K. A. (2002). The content analysis guidebook. Thousand Oaks, Calif: Sage Publications.
- Niekler, A., & Jähnichen, P. (2012). Matching Results of Latent Dirichlet Allocation for Text. In Proceedings of ICCM 2012, 11th International Conference on Cognitive Modeling (S. 317–322). Universitätsverlag der TU Berlin.
- Quasthoff, Uwe / Eckart, Thomas (2009): Corpus Building Process of the Project "Deutscher Wortschatz". In: Linguistic Processing Pipelines Workshop at GSCL 2009.
- Teh, Y. W., & Jordan, M. I. (2010). Hierarchical Bayesian Nonparametric Models with Applications. In N. Hjort, C. Holmes, P. Müller, & S. Walker (Hrsg.), Bayesian Nonparametrics: Principles and Practice. Cambridge University Press.
- Teh, Yee Whye. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In Proceedings of the 21st International Conference on Computational Linguistics, 985–992.

MyCoRe – Eine Software für Repositorien in den Geisteswissenschaften

Wiebke Oeltjen

MyCoRe ist ein Framework zum Erstellen von Repositoryn, Dokumentenservern, Bilddatenbanken und digitalen Archiven. Die Open-Source-Software wird an deutschen Universitätsbibliotheken und -Rechenzentren seit mehr als 10 Jahren kontinuierlich weiterentwickelt. Bundesweit werden an mehr als 20 Standorte über 60 Webanwendungen auf der Basis von MyCoRe betrieben. Der Name [my core] weist darauf hin, dass es sich um einen Softwarekern handelt, der in eigenen, individuell angepassten Webanwendungen eingesetzt werden kann („My“ Content Repository). Der Softwarekern stellt Funktionen bereit, die im Kontext der *Digital Humanities* von Bedeutung sind, weil mit den Webanwendungen digitale Ressourcen und Metadaten erschlossen, archiviert, recherchiert und präsentiert werden können. Sammlungen von Forschungsdaten können ebenso erfasst werden, wie Publikationen oder andere digitale Dokumente (wie z.B. Urkunden, Akten, Handschriften, Bücher, Kataloge, Zeitschriften, Zeitungen etc.), aber auch Bilddateien, Tondokumente und Videos können in MyCoRe-basierten Informationssystemen enthalten sein.

MyCoRe zeichnet sich durch Vielseitigkeit, Anpassbarkeit und Nachhaltigkeit aus:

- Dass MyCoRe **vielseitig** einsetzbar ist, wird durch eine Reihe von MyCoRe-Installationen belegt. Das Spektrum des Einsatzgebietes reicht von Archiven, Bilddatenbanken über Dokumenten- und Zeitschriftenserver bis hin zu Informationssystemen zur Erfassung forschungsrelevanter Daten. Das Framework unterstützt offene Standards und Datenformate, wie z.B. METS/MODS und Dublin Core ebenso wie standardisierte aber auch erweiterbare Klassifikationssysteme. Auch werden gängige Schnittstellen (OAI-PMH etc.) unterstützt, so dass ein Datenaustausch mit verschiedenen Systemen möglich ist. MyCoRe-Anwendungen können für mehrsprachige Nutzung ebenso verwendet werden, wie zur Erfassung der Metadaten in verschiedenen Sprachen. Als Beispiel sei hier die Applikation „Islamische Handschriften“ genannt, die als Online-Bibliothek der Universitätsbibliothek Leipzig arabische, persische und osmanisch-türkische Handschriften digital bereit stellt. Es können digitale Abbildungen verschiedenster Art erfasst und mit integriertem Bildbetrachter präsentiert werden, wie die Informationssysteme „Historische Bestände“ oder „Papyrus und Ostraka Projekt“ belegen.
- MyCoRe-Anwendungen sind **anpassbar** an eigene Anforderungen. Für unterschiedliche Arten zu erfassender Metadaten und Dokumente können Datenmodelle entwickelt bzw. angepasst werden. Auch Klassifikationen können neu erstellt oder mittels Editoren erweitert werden. Eine Rechteverwaltung ermöglicht es, die Zugriffe auf die Daten einzuschränken, falls dies erforderlich ist. Des Weiteren kann ein Rollenkonzept angewendet werden, um Benutzerinnen und Benutzer des Systems zu verwalten. So lassen sich z. B. Benutzungsrechte für Autoren, Editoren oder Redakteure definieren, die in ihren Rollen Daten über Eingabeformulare erfassen, begutachten und veröffentlichen. Die verschiedenen Rollen sind für die Definition eines Arbeitsablaufes (Workflow) in dem System von Bedeutung. Mit solchen Anpassungen ist es nicht nur möglich einfache Archive oder Repositoryn anzulegen, sondern auch virtuelle Forschungsumgebung auszubauen, wie dies z. B. in einem Sonderforschungsbereich an der Fakultät für Geisteswissenschaften der Universität Hamburg geschieht.

- MyCoRe ist **nachhaltig**, weil bewährte Software-Komponenten eingesetzt werden, die als Open-Source-Software zur Verfügung stehen. MyCoRe wird kontinuierlich weiter entwickelt und von bedeutenden Einrichtungen eingesetzt. So betreibt das Statistische Bundesamt eine „Statistische Bibliothek“, in der Publikationen aller Landesvertretungen veröffentlicht werden. Auch setzt ein Bundesministerium die MyCoRe-Anwendung „Open Agrar“ als Publikationsserver ein. Außerdem arbeiten Forschungsprojekte mit MyCoRe-Anwendungen zur Speicherung von Bildmaterial, digitalisierten Handschriften oder anderem Forschungsmaterial. Die Informationssysteme dienen auch zum Erfassen von Forschungsdaten oder zum Veröffentlichen von Forschungsergebnissen. MyCoRe-Repositorien eignen sich außerdem zur langfristigen Speicherung digitaler Daten und Metadaten. Digitale Ressourcen können in den Online-Repositorien mit persistenten Adressen versehen werden, wie z.B. URN, DOI oder Handle, damit sie dauerhaft referenzierbar bleiben.

Mit dem Poster soll gezeigt werden, dass die Software MyCoRe als Basis für Repositorien und Informationssysteme in den Geisteswissenschaften geeignet ist.

Weblinks

- Historische Bestände: <http://archive.thulb.uni-jena.de/hisbest>
- Islamische Handschriften: <http://www.islamic-manuscripts.net>
- MyCoRe-Homepage: <http://www.mycore.de>
- Open Agrar: <https://openagrar.bmely-forschung.de>
- Papyrus und Ostraka Projekt: <http://papyri-leipzig.dl.uni-leipzig.de>
- Statistische Bibliothek der Statistischen Ämter des Bundes und der Länder: <https://www.destatis.de/GPStatistik>

(Alle Webadressen wurden am 20.12.2013 aufgerufen und überprüft)

Literatur

Susanne Dobratz: *Open-Source-Software zur Realisierung von Institutionellen Repositories – Überblick*. In: ZfBB 54 (4-5) 2007, S. 199–206, urn:nbn:de:kobv:11-10081380

Frank Lützenkirchen: *MyCoRe – Ein Open-Source-System zum Aufbau digitaler Bibliotheken*. In: Datenbank Spektrum 2(4), November 2002, S. 23–27

Wiebke Oeltjen: *Virtuelle Bibliotheken flexibel gestalten*. In: Bernhard Mittermaier (Eds.): eLibrary – den Wandel gestalten, Proceedingsband, WissKom 2010, Schriften des Forschungszentrums Jülich, Reihe Bibliothek/Library, Vol. 20, Zentralbibliothek, Verlag, 2010, S. 259–266, <http://hdl.handle.net/2128/3811>

Wiebke Oeltjen: *Vernetzung mit MyCoRe – Eine Repository-Software vernetzt Systeme, Daten und Menschen*. In: Bernhard Mittermaier (Hrsg.): Vernetztes Wissen – Daten, Menschen, Systeme; Proceedingsband, WissKom 2012. Forschungszentrum Jülich GmbH Zentralbibliothek, Verlag, 2012, Seiten 225–233, <http://hdl.handle.net/2128/4699>

MUSICI und MusMig. Kontinuitäten und Diskontinuitäten

Abstract

Musiker waren schon immer eine Berufsgruppe, die eine hohe Mobilität aufwies. Insbesondere in der Frühen Neuzeit sind zahlreiche Musiker bekannt, die aus unterschiedlichsten Gründen ihre Heimat- und Wirkungsorte temporär oder definitiv verließen. Wanderungsbewegungen beschränkten sich dabei nicht auf einzelne Regionen und Länder, sondern erstreckten sich auf ganz Europa und darüber hinaus. Diesem Phänomen und die mit ihm verbundenen Konsequenzen wird aktuell in interdisziplinären und internationalen Forschungsgruppen nachgegangen. Während das Projekt „MUSICI. Musicisti europei a Venezia, Roma e Napoli“ bis 2012 europäische Musiker fokussierte, die zwischen 1650 und 1750 mit verschiedensten Zielsetzungen nach Venedig, Rom und Neapel reisten und dort in unterschiedlichsten Positionen aktiv waren, erforscht das jüngst gestartete Projekt „MusMig. Music Migrations in the Early Modern Age: the Meeting of the European East, West and South“ ab 2013 Migrationsbewegungen von Musikern im 16. und 17. Jahrhundert vor allem im östlichen Europa.

„Migration“ wird dabei nicht nur als „Wanderung“ verstanden, sondern als jede Bewegung im territorialen Raum. Somit fallen auch Reisen von Musikern und Operntruppen zum Zwecke der Aufführung musikalischer Werke, Kavaliersreisen, an denen Musiker teilnahmen, Ausbildungsreisen oder Gesellenwanderungen unter die Untersuchungsobjekte des Projekts. Auch der Begriff „Musiker“ wird weit gefasst und umfasst neben Instrumentalisten, Komponisten und Sängern u.a. auch Textdichter, Instrumentenbauer und Musiktheoretiker. Als Projektergebnis werden Erkenntnisse darüber erwartet, dass die Migration von Musikern maßgeblich zur Dynamik und Synergie der europäischen Kulturszene beigetragen, eine Katalysatorfunktion in Bezug auf Innovationen, auf

stilistische Veränderungen sowie auf den Wandel von musikalischen und sozialen Mustern ausgeübt und insgesamt die Formierung einer gemeinsamen europäischen kulturellen Identität stimuliert hat.

MUSICI und MusMig

Als Fortführung von MUSICI baut MusMig auf den Erfahrungen des Vorgängerprojekts auf – und muss sie angesichts neuer Fragestellungen modifizieren. Während sich z.B. MUSICI auf drei italienische Städte (Venedig, Rom, Neapel) und ihre musikalischen Institutionen konzentrierte und die dorthin reisenden „ausländischen“ Musiker (zu denen auch Musiker der anderen betrachteten Städte zählen mussten, da sie aus einem anderen Territorium stammten) erfasste, ist MusMig breiter angelegt und muss Möglichkeiten bieten, jede Art von Lokalität aufnehmen zu können: vom Hof über die Reichsstadt bis zum ländlichen Kloster. Auf systematischer Ebene hat dies eine stärkere Generalisierung zur Folge, die es ermöglicht, eine größere Bandbreite an Orten zu erfassen. Während für Venedig, Rom und Neapel eine relativ genaue Aufschlüsselung der institutionellen Arbeitgeber von Musikern erfolgen konnte, ist dies aufgrund der Fülle von Institutionen im Untersuchungsgebiet von MusMig nicht mehr möglich. Hier müssen auf genereller Ebene Klassifizierungsmöglichkeiten geschaffen werden, die es dennoch ermöglichen, aussagekräftige Auswertungen zu garantieren. Gleichzeitig müssen angesichts veränderter politischer Rahmenbedingungen etwa Regierungsformen stärker ausdifferenziert werden. Während Venedig eine Republik, Rom eine kirchliche Wahlmonarchie und Neapel ein von einer erbmonarchischen Fremdmacht beherrschtes Territorium war, ist die Bandbreite im Alten Reich und darüber hinaus ungleich größer: vom Heiligen Römischen Reich über Großherzogtümer, Fürstbistümer, Reichsstädte zu weiteren Metropolen (etwa die Stadt Leipzig, die zwar zum Herzogtum Sachsen gehörte, aber aufgrund ihres Status als Messestadt eine Sonderstellung einnimmt).

Die Betrachtung der Systematik soll jedoch nicht im Fokus des Vortrags stehen. Vielmehr wird anhand des bereits abgeschlossenen MUSICI-Projekts dargestellt, wie methodisch mit den Daten auf Visualisierungsebene umgegangen wurde, um dann aufzuzeigen, in welcher Richtung die bestehenden Darstellungsformen weiterentwickelt werden und neue Zugangsmöglichkeiten für Forschende bieten können. Am Beispiel eines Teilprojekts des MusMig-Projekts wird demonstriert, wie sich die Fragestellung durch

digitale Methoden unterstützen lässt und welche Darstellungsformen dafür benötigt werden.

Musikermigration im dynastischen Kontext

Dieses Teilprojekt untersucht ein Konglomerat an Höfen, die dynastisch miteinander verbunden sind und im Zuge der Sukzession miteinander verschmolzen wurden: der Hof der Münchner Wittelsbacher sowie die Wittelsbachischen Nebenlinien Pfalz-Neuburg, Pfalz-Sulzbach und Pfalz-Zweibrücken. Die Untersuchung fokussiert hauptsächlich zwei Aspekte:

1. Inwiefern beeinflussen die dynastische Verbindung einerseits und die notwendigerweise erfolgten Sukzessionen andererseits die Wanderung von Musikern?
2. Inwiefern werden Musiker verschiedener lokaler Abstammung zur höfischen Profilbildung benutzt, wie es allem Anschein nach in Mannheim (böhmische Musiker) und München (italienische Musiker) geschah?

Während im ersten Aspekt vor allem Binnenmigrationen untersucht werden, die durch Residenzverlagerungen ausgelöst wurden, stellt der zweite Aspekt Fragen der Musikerrekrutierung in den Mittelpunkt. Eine Kontextualisierung der Ergebnisse erfolgt durch die Betrachtung von Vergleichshöfen: Hannover (Residenzverlagerung durch Sukzession Georgs von Hannover auf den englischen Thron) und Dresden (dynastische Verschmelzung mit den albertinischen Nebenlinien sowie Profilbildung durch gezielte Musikerrekrutierung).

Die Datengrundlage bilden insbesondere Rechnungs- und Besoldungsbücher, die von den Höfen erhalten sind und aus denen Anstellungszeiten und Gehaltshöhe der Musiker ermittelt werden können. In Einzelfällen finden sich auch weitere Hinweise.

Die digitale Erfassung von Musikern der genannten Höfe mithilfe des sogenannten Aspekt-Datenmodells bietet entscheidende Vorteile. Mit den feinstrukturiert abgelegten Daten lassen sich verschiedenste Auswertungen vornehmen, die einerseits die kurz umrissenen Fragestellungen beantworten können, andererseits ein Fülle von weiteren Informationen bereitstellen, die ein umfassendes Bild des migrierenden höfischen Musikers und der Marktmechanismen in Bezug auf musikalische Human resources im höfischen Umfeld zeichnen. So ist festzustellen, ob und in welchem Maße dynastisch

verbundene Höfe untereinander Musiker kurzfristig oder längerfristig austauschten (siehe Abbildung 1). Ebenso lässt sich eruieren, ob Musiker bereits im Vorfeld einer abzusehenden Sukzession Kontakte zum nachfolgenden Regenten knüpften oder sogar in seine Dienste traten, um ihre Stelle zu sichern (siehe Abbildung 2). Die Untersuchung der Rekrutierung kann Informationen darüber geben, ob es bestimmte Rekrutierungszentren gab, die entweder durch Persönlichkeiten (Gesandte, Residenten, Agenten, andere Vermittler) oder Institutionen (Höfe, kirchliche Institutionen) definiert sind. Sie zeigt außerdem, ob es Musiker gab, die ohne erkennbaren Vermittler eine Anstellung am Hofe fanden. Zusätzlich werden Karrierewege bekannt, die Musiker nahmen, um an einen besonders renommierten Hof zu gelangen.

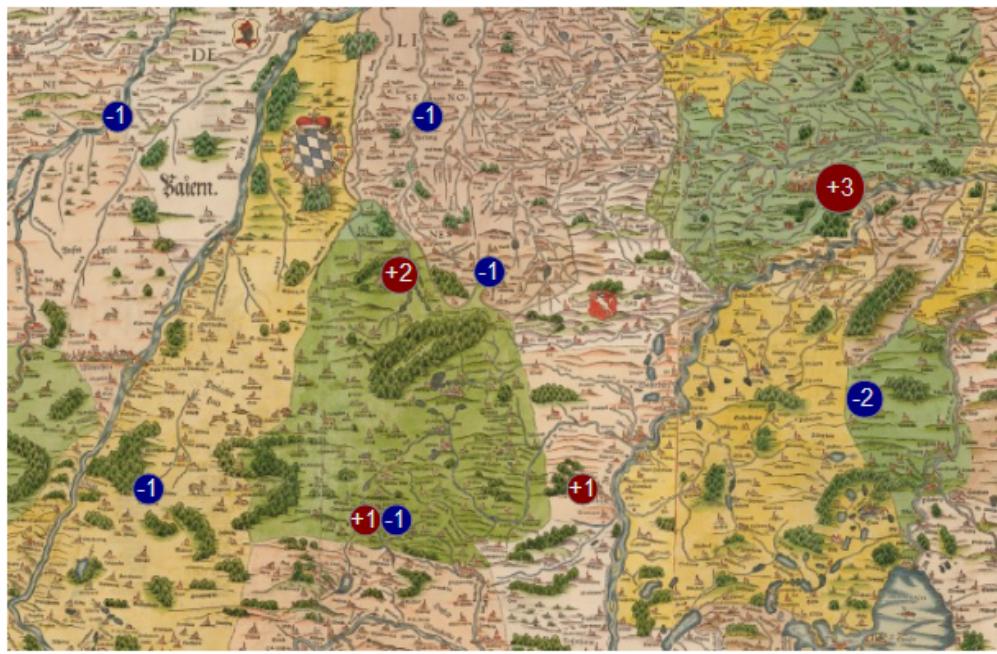
Zahl der erwarteten TeilnehmerInnen

Der Vortrag richtet sich nicht an ein spezifisch musikwissenschaftliches Publikum, sondern an das Konferenzpublikum im Allgemeinen. Im Vordergrund stehen grund-sätzliche Fragen der Auswertung und Visualisierung, die für alle Disziplinen relevant sind. Daher ist mit einer eher hohen Teilnehmerzahl zu rechnen, die selbstverständlich in Relation zur Teilnehmerfrequenz der Konferenz insgesamt steht.

Technische Ausstattung

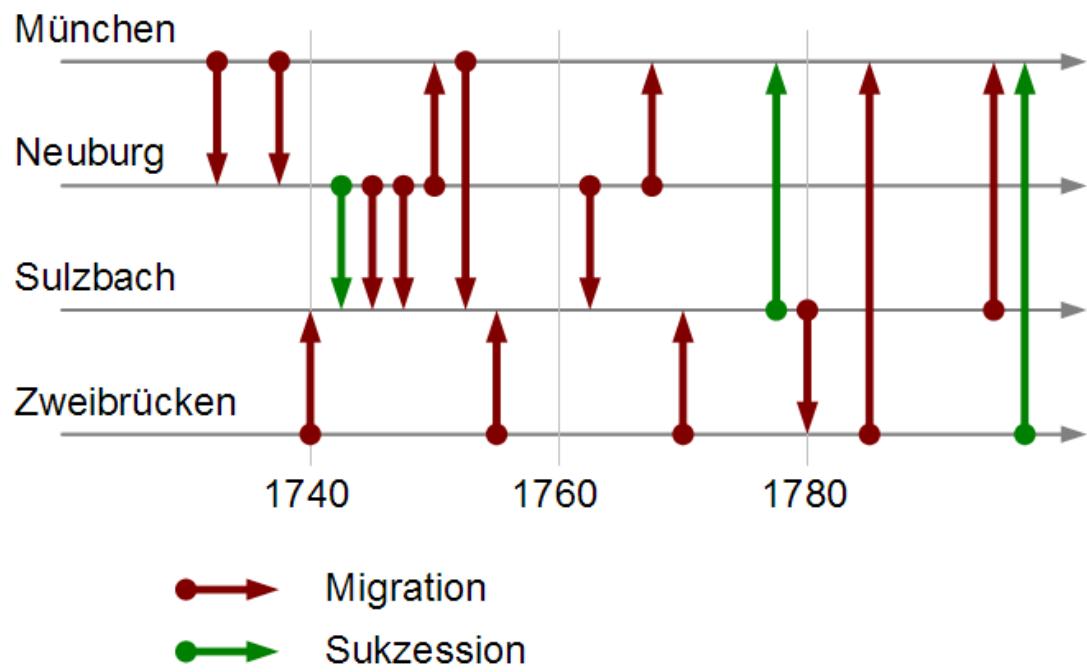
Für den Vortrag wird die konferenzübliche Hardware Laptop und Beamer benötigt.

Abbildung 1



- Rekrutierungsziel

Abbildung 2



Referenten

Dr. Berthold Over

Institut für Kunstgeschichte und Musikwissenschaft – Abteilung Musikwissenschaft
Johannes Gutenberg-Universität
Jakob-Welder-Weg 18
55128 Mainz

Tel +49 / (0)6131 / 39-22781

Mail over@uni-mainz.de

Web <http://www.musikwissenschaft.uni-mainz.de>

Forschungsinteressen

- Georg Friedrich Händel, Antonio Vivaldi
- Musikermigration in der Frühen Neuzeit
- Musik und Aristokratie im 17. und 18. Jahrhundert
- Strategien der höfischen Musikfinanzierung (16.-20. Jahrhundert)

Torsten Roeder

Berlin-Brandenburgische Akademie der Wissenschaften
TELOTA - The Electronic Life of the Academy
Jägerstraße 22/23
10117 Berlin

Tel +49 / (0)30 / 20370-264

Mail roeder@bbaw.de

Web <http://pdr.bbaw.de>

Forschungsinteressen

- Musikwissenschaft
- Italienische Literatur und Linguistik
- Digital Humanities

Abstract für einen Vortrag zum Themenbereich

3) Vom analytischen Mehrwert digitaler Werkzeuge für die Geisteswissenschaften

Der Presenter – Ein digitales Hybridsystem zur Visualisierung und Präsentation von individuellen und kollaborativ erzeugten Sinnstrukturen

In der geisteswissenschaftlichen Auseinandersetzung mit einem Gegenstand gilt es die Verformung der Inhalte durch das genutzte Medium mit zu reflektieren. Wie kann die dabei verwendete technische Unterstützung, z.B. eine Webanwendung, so konzipiert sein, dass sie die Entscheidung über die Gestaltung der Inhalte nicht vollständig übernimmt, sondern die Verantwortung von Genese und Gestaltung von Sinnstrukturen in einem technischen Medium beim Ersteller/bei der Erstellerin liegt?

Im Rahmen des BMBF-Forschungsprojekts *gewiss kühn* (*Gedächtnis wissenschaftlicher Erkenntnisse und künstlerischer Haltungen*)¹ wurde an der HfG Karlsruhe in enger interdisziplinärer Zusammenarbeit von GeisteswissenschaftlerInnen und Informatikern² eine multimediale Lern-, Archiv- und Präsentationsumgebung (*Presenter*) als Hybridsystem entwickelt.³ Die Konzeptformulierung ‚Hybridsystem‘ hebt dabei bewusst den Anspruch hervor, dass das technische System die Absprache und Zusammenarbeit von Menschen fördert und nicht ersetzt. Es ist so angelegt, dass der einzelne Nutzende selbst bestimmt, wie er sich seine gewählten Inhalte aneignet und strukturiert darstellt. Die Open-Source-Webanwendung *Presenter* ist speziell auf den einzelnen Nutzer/die Nutzerin hin konzipiert, Wissen (in Form von Bildern, Volltexten, Audio- und Videodateien etc.) individuell zu strukturieren, für sich multimedial zu visualisieren und zu dokumentieren und an von ihm/ihr bestimmte Schnittstellen zu kommunizieren. Diese Kommunikation wird dahingehend vom System erleichtert, dass die Konzeption und Erarbeitung eines Themas oder Projekts in demselben Medium vollzogen wird wie ihre Präsentation. So vereint er Arbeits- und Präsentationsoberfläche, auf der auch

¹ Projektlaufzeit: September 2011 – August 2014.

² Informatiker, Kunsthistorikerin, Germanistin/Pädagogin, Literaturwissenschaftlerin, Ökonom.

³ Die aktuelle, finale Version des Presenters basiert im Backend auf dem PHP-Framework TYPO3.Flow und auf der Datenbank MySql. Für die Echtzeitkommunikation wird auf nodejs zurückgegriffen, um einen skalierbaren Austausch zu ermöglichen. Das Frontend wurde im Wesentlichen mit jquery-Plugins realisiert, für die Ausgabe von Templates wird auf angularjs zurückgegriffen.

teilfertige Abschnitte von Erarbeitungsprozessen direkt als Präsentation angezeigt werden können.

Die Makro- und Mikrostrukturierung der verschiedenen multimedialen Sinnelemente legt der Nutzer/die Nutzerin im *Presenter* auf einer Arbeitsoberfläche an, d.h. er breitet sie visuell neben-, über- und untereinander aus. So kann die angelegte Sinnstruktur nicht nur durch lineare Rezeption, sondern auch durch das Prinzip der Synchronizität nachvollzogen werden. Wie bei einem komplexen Bildgefüge kann der Rezipient mehrere Elemente gleichzeitig wahrnehmen und auf einen Blick eine Beziehungsstruktur erkennen.

Vor und während der Bearbeitung eines Themas im *Presenter* wird der Nutzer/die Nutzerin vor die Frage gestellt, wie er sich Erkenntnisse im digitalen Raum individuell veranschaulicht: Wie und wann wird für ihn/sie Erkenntnis durch Anordnung und Kombination von multimedialen Inhalten evident?

Innerhalb der digitalen Umgebung besteht darüber hinaus die Möglichkeit, kollaborativ zu arbeiten und so Gruppenwissen zu erzeugen und darzustellen. Durch den Übergang von individuellem zu kollaborativem Wissen wird ein weiterer Reflexionsprozess angestoßen, der nicht allein die Konstruktion des eigenen Wissens beinhaltet, sondern auch die Fragestellung, wie dieses intersubjektiv verfügbar gemacht werden kann. Hier werden Aushandlungsprozesse über die Geltung des Sinnentwurfes in Gang gesetzt. Diese Prozesse werden durch die technisch-systemische Lösung ermöglicht, jedoch nicht übernommen oder automatisiert.

Der *Presenter* bietet Darstellungs- und Vernetzungsmöglichkeiten an, die Lernenden bleiben jedoch Handelnde und immer in der vollen Verantwortung über Auswahl, Strukturierung und Darstellung von Inhalten. Das technische System trifft keine automatisierten Entscheidungen und *erfordert* die Absprache und Zusammenarbeit von Menschen, anstatt sie zu ersetzen.

Inhalte und ihre Struktur bleiben je nach dem aktuellen Wissen des Nutzers bzw. der Nutzergruppe im System immer modifizierbar. Durch eine systemimmanente Versionierungsfunktion werden alle Zustände einer Arbeitsfläche chronologisch archiviert, sodass durch deren Aufrufen die Entwicklung bzw. Veränderung der Wissensdarstellung (und damit auch der Genese von Evidenz) beobachtbar und reflektierbar wird. Dem Nutzer wird somit ermöglicht, evidente Darstellungen als temporär, sinnhafte Strukturen als etwas Prozesshaftes und ‚siche-

re' Erkenntnisse als veränderlich und wandelbar wahrzunehmen. Evidente Darstellungen ‚werden‘, wie man in Analogie zu Joseph Vogls *Medien-Werden* sagen könnte.⁴ Sie entfalten Handlungs- und Wirkmacht und zeugen zugleich von der temporären Aktualität und konventionellen Arbitrarität aller Erkenntnis, allen Sinns und aller Zeichen – bzw. tragen in sich den *Möglichkeitscharakter* des Zeugens und Zeigens.

Diese Konzeption unterscheidet den *Presenter* von den meisten anderen digitalen Lernumgebungen, die von den pragmatischen Zielsetzungen einer Institution ausgehen, die die Kultur des Lernens und die Lerninhalte bestimmt. Die Entwicklung des *Presenters* zielt weniger auf die institutionell vorgegebene Vermittlung von Wissen ab als auf die Stärkung der Kompetenz zum Umgang mit komplexen Informationssachverhalten. Es wird die Möglichkeit geboten, neben der Darstellung von kohärenten (geschlossenen) Aussagesystemen auch eine inkohärente Darstellung von Inhalten zu gestalten, die einen Austauschprozess über das Dargestellte erfordert und die Konstruktion des eigenen Wissens veranschaulicht. Die Offenheit des Dargestellten bedingt die Hinterfragbarkeit, das Verstehen des eigenen Verstehens und die Reflexion der individuellen Produktion von subjektiver Gültigkeit wird forciert.

Das Konzept der Lern- und Archivumgebung *Presenter* bedient sich der Metapher des Gedächtnisses, da es statische Wissensinhalte (organisierte Archive⁵) und dynamische Arbeitsoberflächen vereint, die in einer progressiven Wechselwirkung zueinander stehen. So sind zum einen die eigenen Sinnstrukturen und zum anderen zeitabhängige Umgewichtungen der Informationsbestandteile sichtbar, sodass eine Dokumentation und damit Reflexion des erworbenen und erzeugten Wissens möglich ist. Die entstandenen (Argumentations-)Strukturen von Themenfeldern/Wissenskonvoluten/Texten werden sichtbar. Es entsteht eine Alternative zur linearen Wissensdarstellung und zur hierarchisierten Darstellung von Wissensordnungen.

⁴ Vgl. Joseph Vogl (2001): Medien-Werden. Galileis Fernrohr. In: Mediale Historiographien, 1, S. 115-123.

⁵ Das Archiv des Presenters ist in drei Teile unterteilt: Ein allgemeines Archiv, an das verschiedene Institutionen angeschlossen sind und durch das öffentlich zugängliche Digitalisate verfügbar werden, ein Archiv, das von unterschiedlichen Kollaborationen genutzt werden kann und schließlich ein persönliches Archiv, in dem die benutzen Elemente abgelegt werden. Das kollaborative und das private Archiv können u.a. durch die Quellenfunktion gestaltet werden, die zugleich zwei Vorteile vereint. Zum einen kann durch die selbst getroffene Entscheidung, welche Informationen zu einem Objekt angelegt werden, eine eigene Archivstrukturierung entstehen, zum anderen können die Quellen und Metadaten auf der Arbeitsoberfläche sichtbar gemacht werden, sodass ein wissenschaftlicher Anspruch des Erarbeiteten visuell nachvollziehbar präsentiert werden kann.

Die ‚Verwaltung‘ dieses Wissens erfolgt nicht über eine feststehende Redaktion, sondern ist prozessproduziert von den kollaborativ beteiligten Personen (z.B. WissenschaftlerInnen, Lehrpersonal, VertreterInnen beruflicher Praxis, SchülerInnen/StudentInnen).

Jahrestagung der „Digital Humanities im deutschsprachigen Raum (DHd)“ zum Thema "DH - methodischer Brückenschlag oder 'feindliche Übernahme'? Chancen und Risiken der Begegnung zwischen Geisteswissenschaften und Informatik"

25.-28. März 2014 an der Universität Passau

Abstract (Vortrag)

Zur Sichtbarkeit von Street Art in Flickr. Methodische Reflexionen zur Zusammenarbeit von Soziologie und Informatik

An der Leibniz Universität Hannover wurde eine interdisziplinäre Studie zur Sichtbarkeit von Street Art in Flickr durchgeführt. An der Untersuchung waren ein Soziologe, zwei Informatiker und mehrere Hilfskräfte beteiligt. Ausgangspunkt des Forschungsvorhabens war die kultur- und kunstsoziologische Beobachtung, dass mit dem Phänomen Street Art nicht nur eine Ablösung von der Subkultur des Graffiti Writings, sondern auch ein Wandel der Mediennutzung eingesetzt hat. Einerseits werden heute neben den klassischen Schriftzügen auch andere visuelle Ausdrucksformen verwendet (z.B. Aufkleber, Poster, Schablonengraffiti). Andererseits erfolgt die Kommunikation der Subkultur nicht mehr allein über die Interventionen im öffentlichen Raum, sondern auch über Reproduktionen im Internet.

Innerhalb der Kultur- und Kunstsoziologie hat Wuggenig (2009) die These formuliert, dass sich mit dem Internet auch die Art und Weise der Rezeption und Wahrnehmung von Street Art verändere. Anstatt etablierter Kunstinstitutionen würde vor allem das Internet dazu beitragen, die Sichtbarkeit und Anerkennungsprozesse von Street Artists zu befördern.

In unserer Untersuchung haben wir uns der These angenommen, um exemplarisch für das Internet die Relevanz und die Wahrnehmungsweisen von Street Art am Beispiel des Online-Photoarchivs Flickr zu analysieren. Flickr steht allgemein für das Internet. Zum einen ist das Photoarchiv kein Produkt der Kultur der Street Art (anders bei Internetseiten wie Art Crimes oder Wooster Collective). Zum anderen bietet es eine große Zahl an Reproduktionen von Street Art, um die Art der Wahrnehmung von Street Art und die Rolle von Flickr für das Publikum von Street Art zu untersuchen.

Die Untersuchung setzte die Kooperation von Soziologen und Informatikern voraus, um die erforderlichen Daten zu sammeln und aufzubereiten. Die Analyse bei Flickr konzentrierte sich beispielsweise auf das Jahr 2012 und berücksichtigte über die zufällige Auswahl eines Photos pro NutzerIn eine Grundgesamtheit von 10.868 Bildern. Aus der Grundgesamtheit wurde durch eine einfache Zufallsstichprobenziehung ein Sample von 1.000 fototechnischen Reproduktionen gezogen. Die Auswertung der Bilder erfolgte mit Hilfe der visuellen Inhaltsanalyse. Die Kenntnisse der Informatik ermöglichten, diese Auswahl vorzunehmen, Metadaten zu erheben und die technischen Voraussetzungen für eine Auswertung zu schaffen.

Die technische Umsetzung der Auswahl der Bilder und Metadaten erfolgte mit Hilfe der Informatik durch die Adaptierung eines Flickr Crawlers. Der Crawler fand Verwendung bei der Suche,

Überprüfung der Ergebnisse und Ziehung der Zufallsstichprobe. Es wurde ein Webinterface für die Auswertung und Kodierung der Reproduktionen entwickelt und eingesetzt. Das form-basierte Webinterface zeigte den Kodiererinnen und Kodierern, nach dem Einloggen, die Bilder sowie ein Formular mit den Kodierungen. Des Weiteren wurden die Ergebnisse in einer Datenbank mit Links zu den (auf einem lokalen Webserver) gespeicherten Reproduktionen, deren Metadaten und den Kodierungen zentral gespeichert.

Das Forschungsprojekt kam zu dem Ergebnis, dass in Flickr eine Street Art-"orientierte" Community existiert und für die Sichtbarkeit von Street Art sorgt. Sie sorgen durch determinierende Tags für ein Erkennen und Wiedererkennen von Street Art im Internet. Das Verhältnis dieser informierten Gruppe zur Gesamtheit der Flickr User ist jedoch gering. Folglich zeigt Flickr zwar viele Beispiele für Street Art Objekte, die Internetplattform ist jedoch für die Sichtbarkeit von Street Artists nur von geringer Bedeutung. Es muss jedoch nicht heißen, dass das Internet gar keine Rolle für die Wahrnehmung von Street Art spielt. Die Ergebnisse der Studie legen vielmehr nahe, dass den Internetseiten, die eine enge Bindung an die Kultur der Street Art aufweisen, eine größere Relevanz für die Sichtbarkeit zukommt.

Für die kultur- und kunstsoziologische Studie war von großer Bedeutung, dass sich die Vorgehensweise nach den methodischen Gütekriterien der Soziologie ausrichtet. Der Anspruch leitete sich daher nicht aus der Informatik ab, große Datenmengen zu generieren und ihre Strukturierung zu visualisieren, sondern von soziologischen Theorien und Methoden auszugehen. In der Folge definierten die soziologische Fragestellung und die methodischen Anforderungen das Forschungsvorgehen. Die Informatik übernahm die Bereitstellung einer Dienstleistung für eine soziologisch motivierte Forschung. Dies stellt die interdisziplinäre Zusammenarbeit vor die Schwierigkeit, für beide Seiten forschungsrelevante Daten zu generieren. In unserem Projekt konnte ein Mehrwert für die Informatik aus den soziologisch vorgenommenen Bildkodierungen und zusätzlich erfolgte Klassifizierungen von bildbegleitenden Texten gewonnen werden. Diese Daten konnte wiederum die Informatik für die Erprobung eigener Analyseinstrumente nutzen.

Aus dem Forschungsprojekt ziehen wir daher für die Digital Humanities die vorläufige Schlussfolgerung, dass die Geistes- und Sozialwissenschaften von der Informatik lernen können, aber ebenso die Voraussetzungen brauchen, eigene Forschungsstandards und -themen durchzusetzen. Solche Möglichkeiten bieten sich jedoch nur dort, wo die Informatik hinter ihrem eigenen Forschungsinteresse zurücktritt und in erster Linie als Dienstleistende für die Geistes- und Sozialwissenschaften auftritt. Zugleich muss sichergestellt sein, dass Anschlüsse an die informatikgetriebene Forschung und Rückkopplungen in die Geistes- und Sozialwissenschaften möglich sind.

Die Dariah-DE Architektur zur forschungsorientierten Föderation von Kollektionen in den Digital Humanities

Christoph Plutte¹, Tobias Grndl² und Andreas Henrich²

¹ Berlin-Brandenburgische Akademie der Wissenschaften,
TELOTA und Dariah-DE, Jägerstr. 22/23, D-10117 Berlin

² Universität Bamberg, Lehrstuhl für Medieninformatik,
An der Weberei 5, D-96047 Bamberg

1 Einleitung

Für die kultur- und geisteswissenschaftliche Forschung relevante Ressourcen finden sich zu großen Teilen in den Sammlungen von Museen, Archiven, Bibliotheken, Universitäten und außeruniversitären Forschungseinrichtungen. Mit der Erweiterung des Anwendungsbereiches der Digital Humanities von den Sprachwissenschaften³ hin zu einer ganzheitlichen Sicht auf die Kultur- und Geisteswissenschaften seit den 1990ern [1] wurden vermehrt Methoden, Anwendungen und Standards für die Digitalisierung, Analyse und Beschreibung von Ressourcen geschaffen. Die Menge der heute durch öffentliche Netzwerke verfügbaren und für die kultur- und geisteswissenschaftliche Forschung relevanten Kollektionen steigt nicht zuletzt aufgrund der Verwendung von Zugriffs- und Beschreibungsstandards stetig an und bietet Forscherinnen und Forschern einen potenziellen Zugang zu einer Vielzahl heterogener Ressourcen.

In diesem Vortrag stellen wir eine neuartige Föderationsarchitektur vor, die auf eine Erfassung und Fall-basierte Zusammenführung von Forschungsdaten nach den individuellen Bedürfnissen von Forschungsprojekten abzielt. Digitale Sammlungen werden zentral verzeichnet, zur Vermeidung von Informationsverlusten jedoch nicht harmonisiert, sondern in Form von Beziehungen auf Schemaebene assoziiert, wodurch die Verwendung einer dynamisch föderierten Datenbasis in breiten und interdisziplinären, wie auch in fachspezifischen Anwendungskontexten ermöglicht werden kann [2]. Ein übergeordnetes Ziel besteht insbesondere in der Nutzbarmachung des durch Experten hinterlegten Wissens zu Kollektionen und Daten sowie deren Beziehungen für einen weiten Anwenderkreis.

2 Anwendungskontext

Traditionelle Integrationsansätze folgen häufig dem Muster eines physisch harmonisierten Datenbestands auf Basis eines zentralen Schemas [3,4]. Verteilte und heterogene, semi-strukturierte Daten werden hierbei in ein gemeinsames Schema

³ vgl. die Ausführungen zu *Humanities Computing* in [1]

übersetzt und stehen für eine einfache Weiterverarbeitung in integrierter Form zur Verfügung. Eine zentrale Aufgabe dieses Ansatzes besteht in der Umsetzung eines hinsichtlich der notwendigen Granularität geeigneten Integrationsschemas. In Bezug auf die Digital Humanities als ganzheitliche Anwendungsdomäne, die sich in Form spezifischer, interdisziplinärer und auch übergreifender Informationsbedürfnisse äußert, führt die Integration aller Disziplinen und Perspektiven jedoch entweder zu Schemata kaum verwaltbarer Komplexität oder—bei der Verwendung eines einfachen Modells, wie z. B. Dublin Core—zum Verlust großer Anteile disziplinspezifischer Information.

Für die Konzeption der in Dariah-DE umgesetzten Föderationsarchitektur werden im Folgenden zwei Anwendungsfälle vorgestellt, deren unterschiedliche Anforderungen die Einschränkungen eines solchen zentralistischen Integrationsansatzes verdeutlichen:

Generische Suche Mit der generischen Suche verfolgt Dariah-DE das Ziel, eine übergreifende Suchmöglichkeit zu schaffen, welche die Eigenschaften der Breitens- und Tiefensuche so vereint, dass eine dynamische Anpassung der Suche—z. B. im Hinblick auf eine mögliche Facettierung—erreicht werden kann [5]. Die übergreifende Suche in eng assoziierten Datenquellen erlaubt—unter Anwendung der in der Dariah-DE Crosswalk Registry definierten Assoziationen und Transformationsregeln—eine detaillierte Auseinandersetzung mit den betrachteten Daten (Tiefensuche). Mit einer wachsenden Zahl einbezogener Kollektionen wird die Granularität der Betrachtung und Facettierung ggf. mangels vorhandener Verbindungen reduziert und nimmt die Form einer Breitensuche ein. Für die dynamische Funktionalität der generischen Suche ist die ad-hoc-Integration ausgewählter Kollektionen basierend auf den für eine konkrete Anfrage relevanten Kollektionen und den zwischen diesen vorliegenden Assoziationen erforderlich, um die jeweils zur Verfügung stehende Granularität von Daten nutzen zu können.

Datenintegration Im Gegensatz zu der dynamischen, strukturellen Adaption der generischen Suche an die Zusammensetzung der für eine Anfrage ausgewählten Kollektionen zielen Lösungen der Datenintegration oftmals auf eine Konsolidierung einer a-priori definierten Auswahl von Datenquellen ab [3]. Anforderungen an eine kollektionsübergreifende Integration sind wesentlich von der verfolgten Forschungsfrage abhängig und können z. B. im Kontext der Ablösung von Systemen durch Neuentwicklungen, aber auch für die Ausweitung der Datenbasis einer bestehenden Analyse- und Visualisierungslösung, wie beispielsweise dem Dariah-DE Geobrowser [6], auftreten. Die Anwendung eines zentralen Integrationsschemas bzw. einer zentralen Ontologie führt im Fall der Datenintegration im Gesamtkontext der Digital Humanities zu Problemen, insbesondere wenn eine spezifische Auswahl von Kollektionen für konkrete Forschungsfragen zusammengefasst werden soll. Werden so beispielsweise Kollektionen aus archäologischen und kunsthistorischen Kontexten integriert, so führt die direkte Integration der spezifischen Datenstrukturen zu einem erhöhten Informationsgehalt gegenüber einer globalen Struktur, die den Fachspezifika nicht gerecht werden kann.

3 Föderationsarchitektur

Die in Dariah-DE gewählte Architektur (vgl. Abbildung 1) besteht aus der *Collection Registry* zur Verzeichnung von Kollektionen, der *Schema Registry* zur Verwaltung von Schemata, und der *Crosswalk Registry* zur Beschreibung von Assoziationen zwischen verschiedenen Schemata. Integrative Dienste wie die *generische Suche* setzen für die Interpretation und Verarbeitung von Daten der verzeichneten Kollektionen auf den durch die Registries angebotenen Webervices auf.

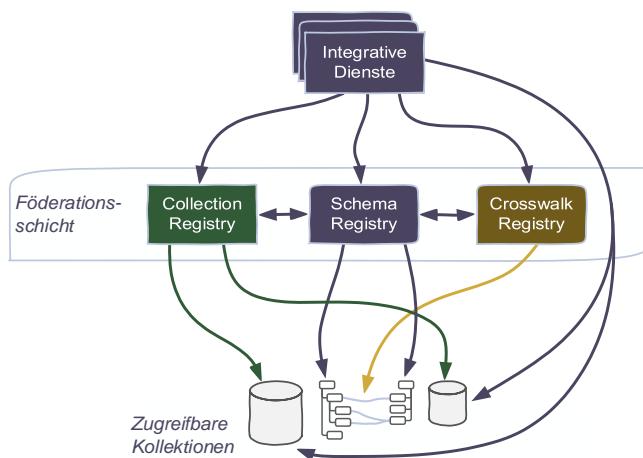


Abb. 1. Komponenten und Zusammenwirken der Föderationsarchitektur

Für eine Forscherin, die eine Sammlung im Rahmen der Föderationsarchitektur registrieren und damit für die Suche, Analyse und den Vergleich mit anderen Sammlungen zur Verfügung stellen möchte, ergibt sich im Zusammenspiel mit der generischen Suche ein Ablauf in vier Schritten (vgl. Abbildung 2):

1. Wenn die entsprechende Kollektion noch nicht in der Collection Registry verzeichnet ist, wird in einem ersten Schritt eine neue Beschreibung der Kollektion und insbesondere ihrer Zugriffsdienste angelegt.
2. Im zweiten Schritt kann die Forscherin das in der Kollektion verwendete Schema (Dublin Core, Lido etc.) beschreiben bzw. die konkrete Verwendung eines allgemeinen Schemas spezifizieren. (Vererbung)
3. Das so erstellte bzw. angepasste Schema kann in Abhängigkeit von konkreten Forschungsfragen im dritten Schritt iterativ mit weiteren Schemata assoziiert werden. (Definition von Crosswalks)
4. Im vierten Schritt indiziert die generische Suche die zugreifbaren Daten der Kollektion anhand der in den Registries hinterlegten Informationen und stellt diese für übergreifende Suchanfragen bereit.

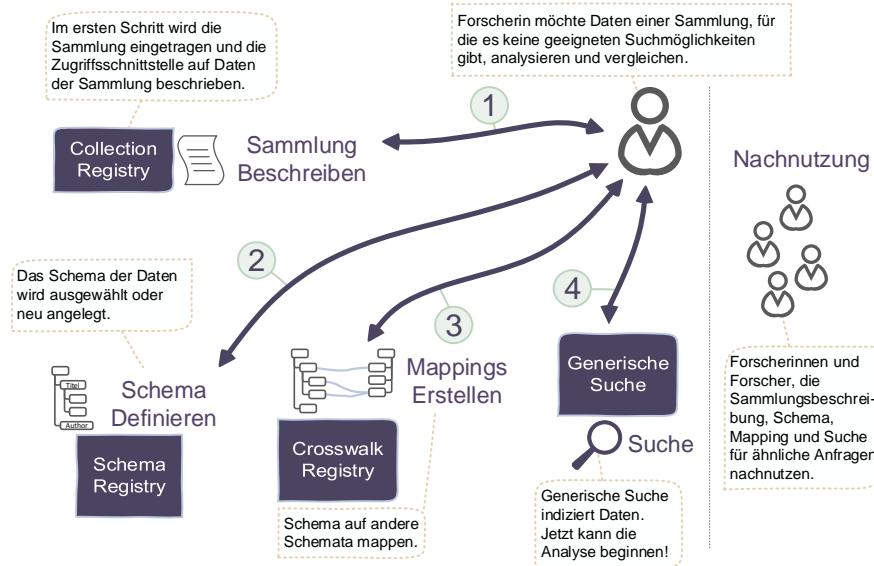


Abb. 2. Schritte der Registrierung von Kollektionen und Schemata

Die sich aus den einzelnen Schritten ergebenden Informationen stehen zur Nachnutzung für verwandte Forschungsinteressen zur Verfügung und können von integrativen Diensten über Webservices abgefragt werden.

3.1 Collection Registry

Die Collection Registry⁴ ist ein online zugängliches zentrales Verzeichnis, in dem relevante Sammlungen registriert und durch Fachwissenschaftler beschrieben werden. Das Datenmodell für die Sammlungsbeschreibungen basiert auf dem Dublin Core Collection Application Profile [7], das insbesondere im Hinblick auf die Beschreibung von Zugriffspunkten erweitert wurde. Die Sammlungsbeschreibungen decken neben Verschlagwortung, zeitlichen und geografischen Dimensionen auch Sammlungsformate und Informationen zur Datenpflege ab. Ein Schwerpunkt liegt auf der Beschreibung von Zugriffspunkten wie OAI-PMH-Schnittstellen zur Abfrage der Sammlungselemente für die Weiterverarbeitung durch assoziierte Komponenten. Komponenten können alle erforderlichen Informationen für einen Zugriff auf die Sammlungselemente aus der Collection Registry über Webschnittstellen (REST) beziehen [8].

Neben maschinenlesbaren Schnittstellen für den Zugriff auf die Sammlungsbeschreibungen bietet die Collection Registry ein Benutzerinterface, welches das Anlegen von Sammlungsbeschreibungen und anderen Datenobjekten ebenso unterstützt wie das Suchen, Aktualisieren und Löschen von vorhandenen Beschreibungen. Ausgewählte kontrollierte Vokabulare unterstützen die Eingabe und die

⁴ <http://demo2.dariah.eu/colreg/>

Interaktion mit der Schema Registry erlaubt es, eine Sammlungsbeschreibung mit einem bestimmten Schema zu verknüpfen. Für den langfristigen Betrieb wird eine Moderation von Dariah-DE organisiert, die die Qualität der Daten gewährleisten wird.

3.2 Schema- und Crosswalk Registry

In der Schema- und Crosswalk Registry⁵ werden semi-strukturierte Datenmodelle und Korrelationen (siehe Abbildung 3) zwischen diesen aus der primären Zielsetzung heraus beschrieben, expliziertes Expertenwissen zu Kollektionen und den darin verwalteten Daten nachnutzen zu können. Die Spezifikationen von Strukturen z. B. in XML Schema können hierbei in Bezug auf eine Kollektion erweitert und konkretisiert werden, wodurch die Semantik originärer Daten erhalten bleibt und dennoch eine Verfeinerung um zunächst implizites Hintergrundwissen erfolgen kann.

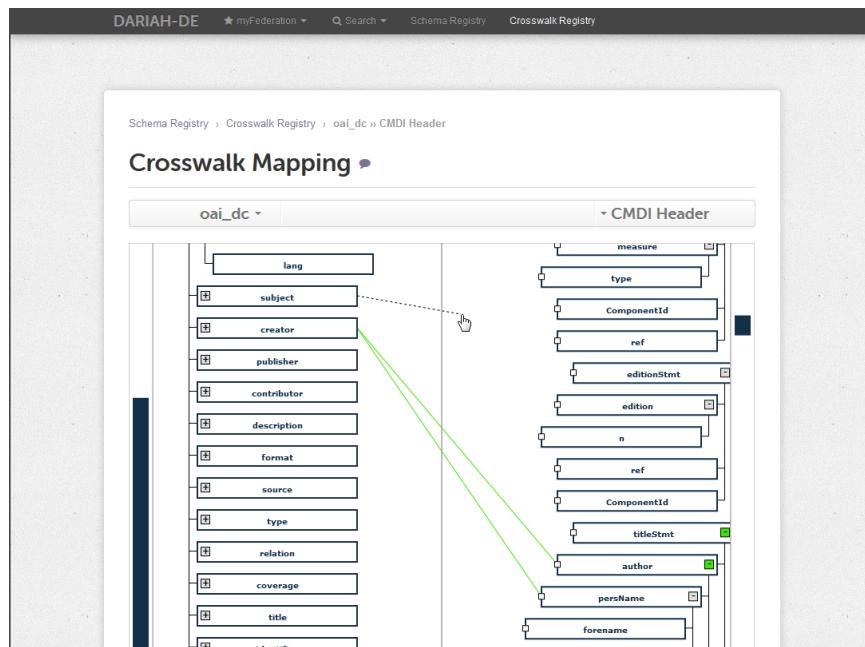


Abb. 3. Assoziation von Schemata in der Crosswalk Registry

Abbildung 4 zeigt beispielhaft Möglichkeiten zur Verfeinerung von Dublin Core basierend auf dem Wissen zu spezifischen Kollektionen. Manuell modellierte Verarbeitungsregeln führen dabei zu einer erweiterten Version eines Datensatzes, welcher für ein Mapping mit komplexeren Strukturen zur Verfügung steht.

⁵ <http://dev3.dariah.eu/schereg/>

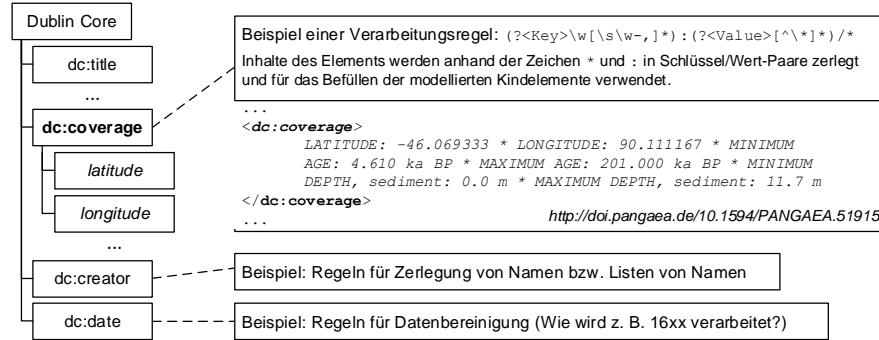


Abb. 4. Beispiele zur kollektionsspezifischen Ergänzung von Dublin Core

Dadurch, dass auch der unveränderte Datensatz weiterhin verwendet werden kann, wird zudem die Kompatibilität zu generischem Dublin Core sichergestellt.

3.3 Generische Suche als durchgeföhrter Use-Case

Mit der generischen Suche⁶ wird im Rahmen von Dariah-DE ein Anwendungsfall der Datenföderation umgesetzt. Hierbei werden Daten aus den in der Collection Registry verzeichneten Kollektionen nach den in der Schema Registry explizierten Strukturen verarbeitet und indexiert. Die Heterogenität der Ressourcen wird zum Zeitpunkt konkreter Suchanfragen basierend auf der zu durchsuchenden Menge von Kollektionen mit Hilfe der Crosswalk Registry aufgelöst.

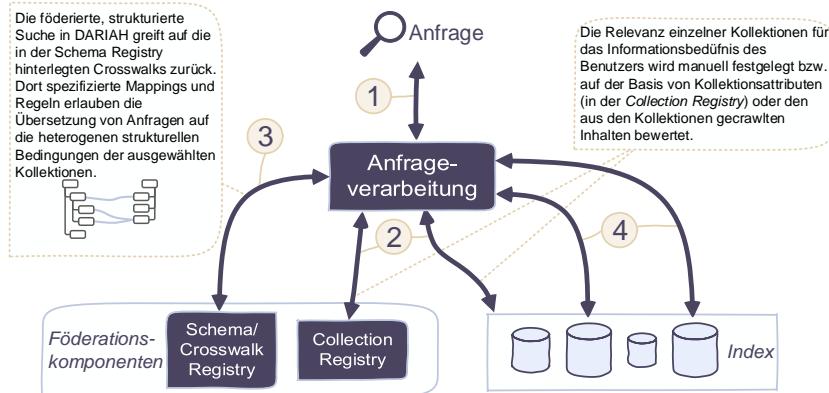


Abb. 5. Anfrageverarbeitung in der generischen Suche

⁶ <http://dev3.dariah.eu/search/>

Abbildung 5 skizziert den Verlauf der Anfrageverarbeitung und die Interaktion mit den Komponenten der Föderationsarchitektur: Am Beginn steht ein Informationsbedürfnis im Rahmen einer Forschungsfrage (1). Zunächst wird nun interaktiv oder automatisch auf Basis der Collection Registry und der von der generischen Suche angebotenen Kollektionssuche die Teilmenge der Kollektionen ermittelt, auf denen die Suche durchgeführt werden soll (2). Je feingranularer die Schemata der gewählten Kollektionen in der Crosswalk Registry miteinander verknüpft sind, umso differenzierter können die Anfragen spezifiziert und ausgeführt werden. Der Nutzer kann die Anfrage dabei in einem Schema seiner Wahl formulieren, das als temporäres Integrationsmodell genutzt wird. Die Anfrage wird auf Basis der relevanten Schemainformationen und Transformationsregeln (3) dann so transformiert, dass sie auf den Indices, die die Daten in ihrem ursprünglichen Schema verwalten, ausgeführt werden kann (4). Ermittelte Ergebnisse werden zusammengefasst und bzgl. ihrer Relevanz für die Anfrage sortiert.

4 Zusammenfassung

Die vorgestellte Föderationsarchitektur folgt dem Prinzip der dezentralen Integration von Daten. Mit der generischen Suche kann gezeigt werden, wie durch die Verwendung der einzelnen Föderationskomponenten ein echter Mehrwert für die Recherche über verschiedene heterogene Datensammlungen entstehen kann und wie eine Alternative zu zentralistischen Ansätzen entwickelt werden kann. Mit einer ad-hoc Föderation kann gegenüber einer domänenweiten Harmonisierung die Möglichkeit der individuellen Integrierbarkeit von Daten geschaffen werden, die auf dem Wissen und der Kollaboration von Spezialisten aus verschiedenen Fachwissenschaften basiert und durch ein breites Publikum in Abhängigkeit von konkreten Forschungsfragen konkret eingesetzt werden kann.

Literatur

1. S. Schreibman, R. G. Siemens, and J. Unsworth, Eds., *A companion to digital humanities*, ser. Blackwell companions to literature and culture. Malden and Mass: Blackwell Pub., 2004, vol. 26.
2. A. Henrich and T. Grndl, “DARIAH(-DE): Digital Research Infrastructure for the Arts and Humanities — Concepts and Perspectives,” *International Journal of Humanities and Arts Computing*, vol. 7, no. supplement, pp. 47–58, 2013.
3. M. Lenzerini, “Data Integration: A Theoretical Perspective,” in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, S. Abiteboul, Ed. New York and NY: ACM, 2002, p. 233.
4. S. Peroni, F. Tomasi, and F. Vitali, “Reflecting on the Europeana Data Model,” in *Digital Libraries and Archives*, ser. Communications in Computer and Information Science, M. Agosti, F. Esposito, S. Ferilli, and N. Ferro, Eds. Berlin and Heidelberg: Springer Berlin Heidelberg, 2013, vol. 354, pp. 228–240.
5. T. Grndl and A. Henrich, “DARIAH-DE Generische Suche (M 1.4.2.1 - Prototyp): DARIAH-DE Arbeitspapier,” 2013. [Online]. Available: <https://dev2.dariah.eu/wiki/download/attachments/2295542/Report%20M1.4.2.1.docx>

6. M. Romanello, “DARIAH Geo-browser: Exploring Data through Time and Space,” 2013. [Online]. Available: <http://de.slideshare.net/56k/dariah-geobrowser-exploring-data-through-time-and-space>
7. Dublin Core Metadata Initiative, “Dublin Core Collections Application Profile,” 2007. [Online]. Available: <http://dublincore.org/groups/collections/collection-application-profile/>
8. C. Plutte and P. Harms, “Collection Registry (M 1.2.2): DARIAH-DE Arbeitspapier,” 2012. [Online]. Available: https://dev2.dariah.eu/wiki/download/attachments/14651583/M1.2.2_Collection_Registry_incl_DCLAP.pdf

Unter Rubrik „Beispiele für disziplinspezifische Anwendungen in der ganzen Breite der Geisteswissenschaften, sowohl in ihren objektbezogenen (Archäologie, Ur- und Frühgeschichte, Kunstgeschichte etc.) als auch in ihren textbezogenen Ausprägungen.“

SlaVaComp – Kirchenslavisch digital: wozu?*

Bei dem Vortrag handelt es sich um einen Erfahrungsbericht aus dem Projekt SlaVaComp, das derzeit als ein Kooperationsprojekt zwischen dem Slavischen Seminar und dem Rechenzentrum der Universität Freiburg durchgeführt wird. Ziel des Projekts ist es, aus mehreren unterschiedlich formatierten griechisch-kirchenslavischen bzw. kirchenslavisch-griechischen Glossaren ein zweisprachiges Metaglossar zu erstellen, das die lexikalische Variabilität des Kirchenslavischen in seiner regionalen und chronologischen Entwicklung erfasst.

Bereits in den ersten Monaten der Projektarbeit ergaben sich mehrere Probleme philologischer, linguistischer und informatischer Art, die im Vortrag zur Diskussion gestellt werden sollen. Aus der Sicht der Philologie wird dabei der Schwerpunkt auf die Frage gelegt, inwieweit uns die moderne Computertechnologie hilft, unser Wissen über die erste Schriftsprache der orthodoxen Slaven auf einem qualitativ neuen Niveau darzustellen und strittige Fragen zu beantworten sowie Desiderata der historischen Slavistik zu erfüllen. Aus der Sicht der Informatik soll auf die Lösungen eingegangen werden, mit denen linguistisch und technisch heterogene Ausgangsdaten in den Quelldateien in eine komplexe Datenbank mit einem mehrstufigen System von Auswahl- und Suchoptionen vereinigt werden.

Die lexikographische Erfassung des Kirchenslavischen gehört zu den Problemfeldern der slavischen Philologie. Wir verfügen bis heute nur über ein Wörterbuch, das dieses Idiom in seiner ältesten Entwicklungsstufe darstellt. Es handelt sich dabei um das sog. Prager Wörterbuch des Altkirchenslavischen. Es umfasst den Wortschatz der Kanontexte aus dem 10. – 11. Jh., d. h. der Texte, auf deren Grundlage das Altkirchenslavische als Sprache rekonstruiert wurde. Eine weitere Entwicklung dieses Idioms in unterschiedlichen Regionen der Slavia orthodoxa vom 11. bis 17. Jh. bleibt immer noch ohne lexikographische Erfassung, die modernen Anforderungen der Sprachgeschichtsforschung entsprechen würde. Dies führt zu Fehlinterpretationen in slavischen Nationalphilologien.

In den letzten zwanzig Jahren wurden von makedonischen, bulgarischen, serbischen und russischen Philologen mehrere Wörterverzeichnisse zu den Editionen slavischer mittelalterlicher Texte vorbereitet. Diese in Papierform vorhandenen Glossare ergänzen zwar wesentlich unsere Vorstellungen über die Entwicklung des Kirchenslavischen im Laufe der sieben Jahrhunderte. Jedoch erlauben sie nicht, diese Entwicklung als ein vollständiges Bild darzustellen. Letzteres ist erst dann möglich, wenn alle Glossare zu einem Metaglossar zusammengeführt würden. Gerade dies erlaubt das digitale Format. Mehr noch: Ein digitales Metaglossar ermöglicht uns, die regionale und funktionale Heterogenität des Kirchenslavischen, darunter auch in den Kanontexten, auf denen das Prager Wörterbuch basiert, of-

* Das Projekt „SlaVaComp – COMPutergestützte Untersuchung von VAriabilität im KirchenSLAvischen“ wird vom BMBF gefördert (FKZ: 01UG1251, Laufzeit: 15.01.2013–15.01.2016). Dr. Irina Podtergera ist Stipendiatin des Margarete von Wrangel-Habitationsprogramms für Frauen (gefördert durch das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg).

fensichtlich zu machen. Dies widerlegt die immer noch gängige Meinung über das Kirchenslavische als ein in sich geschossenes System.

Im Vortrag soll illustriert werden, wie ein digitales Wörterbuch des Kirchenslavischen traditionelle Vorstellungen über diese Sprache ändert und von welcher Bedeutung dies für die weitere Erforschung der slavischen Sprachgeschichte ist. Daneben sollen linguistische Probleme zur Sprache kommen, die in den Altphilologien häufig unbeachtet bleiben, was eine negative Auswirkung auf die Darstellung des sprachlichen Stoffs und seine darauf folgende wissenschaftliche Interpretation hat. Es existiert beispielsweise kein Wörterbuch des Mittelgriechischen. Die Wörterbücher, die in der klassischen Philologie benutzt werden, sind in erster Linie Wörterbücher des klassischen Griechischen. Übersetzungen ins Kirchenslavische wurden jedoch aus dem christlichen Griechischen gemacht. Es gibt zwar Lexika, die das patristische oder das Bibelgriechische fixieren (Lampe, Bauer, Muraoka). Es gibt jedoch keinen Standard, nach dem das Mittelgriechisch in den Wörterbüchern dargestellt werden soll. Auch das „Lexikon zur byzantinischen Gräzität“, bei dem es sich im Grunde genommen um eine Ergänzung zum Liddell/Scott-Wörterbuch des klassischen Griechischen handelt, erfasst nicht das Mittelgriechische als Ganzes. Dieses Problem ist aus der Sicht der Paläoslavistik alles andere als trivial, weil sich mittelalterliche slavische Schreiber bei Übersetzungen aus dem Griechischen des Wortgebrauchs ihrer griechischen Quellen bedienten. Das von uns erstellte Metaglossar macht diese Diskrepanz offensichtlich. Dies soll im Rahmen des Vortrags an konkreten Beispielen erörtert werden. Ein besonderer Akzent wird dabei auf das Problem der Lemmatisierung des Mittelgriechischen für die Bedürfnisse der Altslavistik gelegt werden.

Einen weiteren Schwerpunkt des Vortrags bildet die Frage, wie philologische und historisch-sprachwissenschaftliche Probleme mit Hilfe der Informatik behoben werden können.

Auf technischer Ebene ergeben sich auf dem Weg von den einzelnen Glossaren zum Metaglossar mehrere Schwierigkeiten. Zunächst müssen die Glossare normalisiert werden. Obwohl alle Glossare, die in der Datenbank enthalten sein werden, bereits in digitaler Form (Microsoft-Word-Dateien) vorliegen, bedürfen sie einiger Nachbearbeitung. Fast alle Dateien sind zu einem Zeitpunkt entstanden, bevor die darin dargestellten Buchstaben Teil des Unicode-Standards waren. Da es noch keinen Font gab, der alle diese Buchstaben hätte abbilden können, wurden verschiedene Schriftarten für die Abbildung verwendet. Eine einfache Portierung in das neuere .docx-Format unter Verwendung einer einzigen Unicode-Schriftart wird dadurch unmöglich. Diese Schwierigkeit wurde gelöst durch die Programmierung einer kleinen Anwendung, die alle in einem Dokument enthaltenen Zeichen von Nicht-Unicode-Fonten in Unicode-Zeichen umwandelt und in einem Unicode-Font, d. h. Roman Cyrillic, darstellt. Der nächste Schritt sieht die Transformation der Worddateien in TEI-konformes XML vor. Auch das ist nicht trivial, weil die Glossare sehr heterogene Strukturen aufweisen, was bei der Strukturierung der XML-Dateien berücksichtigt und entsprechend umgesetzt werden muss.

Ein weiteres größeres Problem stellt der Umgang mit der graphischen Variabilität dar. In der XML-Datei wird jedem Lemma ein Hyperlemma, d. h. eine standardisierte Form zugewiesen. Während dieser Vorgang für das Griechische relativ einfach ist, weil es orthographische Standards gibt, nach denen man automatisiert Hyperlemmata erzeugen kann (z.B. bei der Erzeugung der 1. Person von Verben), muss dies fürs Kirchenslavische, für das die Hyperlemmata dem Standard des Prager Wörterbuchs folgen, (noch) weitgehend manuell erledigt werden. Derzeit wird an einer Möglichkeit gearbeitet, diese Normen automatisiert zu generieren, was jedoch sehr zeitaufwendig ist, da aufgrund der Formenvielfalt für fast jede graphische Realisierung eine eigene Transformationsregel formuliert werden muss.

Sobald alle Glossare fertig in XML-Dateien umgewandelt und die Probleme bezüglich graphischer Variabilität behoben sein werden, kann mit der Erschließung der Datenbank begonnen werden. Dazu ist ein Webservice mit einer je nach Anspruch des Benutzers mehr oder weniger ausführlichen Such- und

Filterfunktion in Vorbereitung. Mit Hilfe dieses Webservices kann das Kirchenslavische schließlich erstmals in seiner ganzen regionalen und funktionalen Heterogenität erfasst und beschrieben werden.

Somit wird deutlich, von welcher Relevanz das digitale Wörterbuchformat mit seiner dynamischen und flexiblen Struktur, die im traditionellen Papierformat ausgeschlossen ist, für die Erschließung der Idiome mit stark ausgebildeter graphischer, morphologischer u. a. Variabilität ist. Darüber hinaus ist es offensichtlich, dass die Lösung einer solchen Aufgabe nur im Rahmen einer interdisziplinären Kooperationsarbeit zwischen Philologen und Informatikern möglich ist.

Theodor Fontanes Notizbücher.

Konzept und Beispiele der genetisch-kritischen und kommentierten Hybrid-Edition

Notizbuch-Editionen erfordern komplexe philologische Methoden der Transkription, der Textkonstitution und der Kommentierung, die sich an der Materialität und Medialität sowie an den Funktionen und Inhalten von Notizbüchern orientieren müssen. Der Materialität – dem Format, dem Nach- und Nebeneinander beschrifteter und unbeschrifteter Blätter, den Blattfragmenten und aufgeklebten Blättern sowie den Schreiberhänden, dem Schreibwerkzeug und -duktus – kommt dabei eine besondere Bedeutung zu, da die Analyse der materialen Beschaffenheit eine wichtige Grundlage für funktionale und inhaltliche Aussagen über Notizbücher bildet. Damit die Informationen über die Materialität des Überlieferungsträgers in einer Edition nicht verloren gehen, ist die angemessene Wiedergabe materialer Kennzeichen entscheidend für das editorische Konzept.

Konventionelle Verfahren, die sich lediglich auf die inhaltliche Wiedergabe der Notate konzentrieren und diese wohlgeordnet in einer Buchedition veröffentlichen, bieten nicht nur sehr begrenzte Möglichkeiten, die vielfältigen medialen, funktionalen und materialen Eigenschaften von Notizbüchern aufzubereiten. Sie verhindern zudem, dass alle substantiellen Merkmale, die sowohl die Notizbücher in ihrer einmaligen physischen Gestalt charakterisieren als auch ihre Einträge maßgeblich beeinflussen, von den Lesern und Benutzern rezipiert und für Forschungszwecke ausgewertet werden können. Zwar wurden innerhalb der neugermanistischen Editionswissenschaft die Transkriptionsprinzipien und -verfahren seit D. E. Sattlers Frankfurter Hölderlin-Ausgabe immer weiter ausdifferenziert und graphisch umgesetzt, aber erst die digitalen Methoden, die seit nunmehr zwei Jahrzehnten die Editionsphilologie bereichern, ermöglichen eine annähernd zeichen- und positionsgetreue Auszeichnung, Codierung und Darstellung sowie die Bereitstellung der Transkriptionsdaten zur rechnergestützten Analyse und Weiterverarbeitung.

Am Beispiel der genetisch-kritischen und kommentierten Hybrid-Edition von Theodor Fontanes Notizbüchern soll die Notwendigkeit und der Mehrwert digitaler Editionen demonstriert werden. In der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz sind insgesamt 67 Kleinoktavbändchen überliefert, in denen Fontane zwischen 1860 und Ende der 1880er Jahre poetische Entwürfe, journalistische Aufzeichnungen, Exzerpte, Tagebuchniederschriften und Briefkonzepte niedergeschrieben hat, aber auch seine Alltagsnotizen wie Kochrezepte, Vokabellisten, Zugabfahrtspläne und To-do-Listen festhielt. Hinzu kommen zahlreiche Skizzen von Sehenswürdigkeiten, Grabmonumenten und

Schlachtplänen, die er während seiner Exkursionen durch die Mark Brandenburg und der Reisen nach Dänemark, Böhmen, Frankreich und innerhalb Deutschlands angefertigt hat. Fontanes Notizbücher sind bisher von der Literatur- und Kulturwissenschaft nicht rezipiert worden, obwohl sie gerade durch ihre heterogenen Notate werkgenetische und biographische Informationen enthalten und mehr als andere Handschriftenkonvolute Einblicke in Fontanes schriftstellerische Arbeitsweise ermöglichen. Die Ursachen für die ausgebliebene Rezeption der Notizbücher Fontanes sind vielschichtig; die Hauptverantwortung trägt die vorherrschende Editionspraxis. So wurden bisher nur kleinere Teilveröffentlichungen in gedruckten Medien vorgelegt, in denen die Auswahl der Niederschriften allein nach inhaltlichen Kriterien festgelegt wurde. Die Materialität und die Medialität der Notizbücher Fontanes sind dabei weitgehend unberücksichtigt geblieben, was zu literaturwissenschaftlichen Fehleinschätzungen geführt hat.

Der Vortrag wird die im Notizbuch-Projekt angewendeten digitalen Methoden und den interdisziplinären Workflow von Editions- und Informationswissenschaft vorstellen sowie Einblicke in die Arbeit mit der Virtuellen Forschungsumgebung TextGrid geben, innerhalb der nicht nur die Codierung erfolgt, sondern vielmehr auch Visualisierungen generiert werden, die ein hohes Maß an Auszeichnungsgenauigkeit und -tiefe ermöglichen: das digitale Pendant zur Komplexität der realen Notizbücher.

Als Auszeichnungssprache wird hierbei XML nach den aktuellen Regeln der Text Encoding Initiative (TEI) verwendet. TEI in der aktuellen Version P5 hat sich zu *dem* De-facto-Standard für die Codierung digitaler Editionen entwickelt. Dabei ist das Regelwerk geradezu berücksichtigt für seine Komplexität. Wie das Beispiel der Notizbuch-Edition zeigt, sind weitergehende TEI-Standardisierungen (z. B. TEI Lite, DTA-Basisformat) oder TEI-Anpassungen aus ähnlichen Editionsprojekten nicht in jedem Fall nutzbar, da nicht nur Unterschiede im Quellenmaterial, sondern vor allem unterschiedliche Zielsetzungen bei der Erschließung desselben projektspezifische Richtlinien zur TEI-Anwendung erforderlich machen. Der TEI-Code der Fontane-Notizbücher basiert im Wesentlichen auf der durch das TEI-Modul „transcr - Representation of Primary Sources“ vorgegebenen Grundstruktur, die besonders für (ultra-)diplomatische Transkriptionen und genetische Editionen geeignet ist, und kombiniert diese mit Elementen aus anderen Modulen wie zum Beispiel „namesdates – Names, Dates, People, and Places“ zur inhaltlichen Erschließung von referenzierten Entitäten. Das Ergebnis ist ein Codierungsschema, das TEI-konform bleibt und zugleich den vielfältigen editionsphilologischen Anforderungen entspricht.

Zudem werden im Rahmen eines an die TextGrid-Infrastruktur angeschlossenen Projektportals neben der Edition als (ultra-)diplomatische Transkription, dem Edierten Text und der Präsentation der Faksimiles sowie verschiedene Kommentare auch Datenaggregationen und erweiterte Suchfunktionen angeboten, die unter Einbeziehung von Linked Open Data die Darstellungsmöglichkeiten der Edition erweitern werden. Dabei handelt es sich um eine eXist-Datenbank, die ein im Kontext des SADE-Projektes entwickeltes Modul zur Übertragung und Darstellung von Daten aus dem TextGrid-Repository nutzt.

Zu den Visualisierungsmöglichkeiten gehört auch die Nutzung des Dariah Geo-Browsers, welcher basierend auf den durch den philologischen Kommentar verifizierten Ortsentitäten und zugehörigen, im Umfeld genannten Kalenderdaten in den Notizbüchern eine Raum-Zeit-Relation präsentiert. Hinzu kommt die Generierung von Kookkurrenz-Netzwerken mit Hilfe der D3.js Visualisierungsumgebung. Hierzu wird das TEI-Element <rs> ausgewertet und es werden Zusammenhänge auf der Grundlage eines Notizbuches, einer Seite oder eines Absatzes dargestellt.

Die genannten technischen Innovationen gehören nicht primär zur Edition im klassischen Sinne. Vielmehr sind sie Grundlage der sicheren und ortsunabhängigen kollaborativen editionswissenschaftlichen Arbeit und der Forschungsdatenerarbeitung im Allgemeinen sowie ein Mittel zur detailgetreuen Darstellung. Sie erweitern und bereichern in diesem Projektkontext die praktische Editionsarbeit, ohne die editionswissenschaftlichen Prinzipien, Methoden und Arbeitsweisen zu beschränken oder neu zu definieren. Sie bilden vielmehr die entscheidende Voraussetzung dafür, dass Fontanes Notizbücher, die bisher als unedierbar galten, nach dem überlieferungsadäquaten editionsphilologischen Prinzip der Materialität veröffentlicht und für die weiterführende wissenschaftliche Arbeit zur Verfügung gestellt werden können.

Die interdisziplinär erarbeitete Edition entsteht an der Theodor Fontane-Arbeitsstelle der Universität Göttingen in enger Zusammenarbeit mit der Niedersächsischen Staats- und Universitätsbibliothek Göttingen (SUB) und wird von der Deutschen Forschungsgemeinschaft seit Juni 2011 gefördert.

Dr. Gabriele Radecke, Theodor Fontane-Arbeitsstelle Göttingen,
Martin de la Iglesia (SUB Göttingen) und Mathias Göbel (SUB Göttingen)

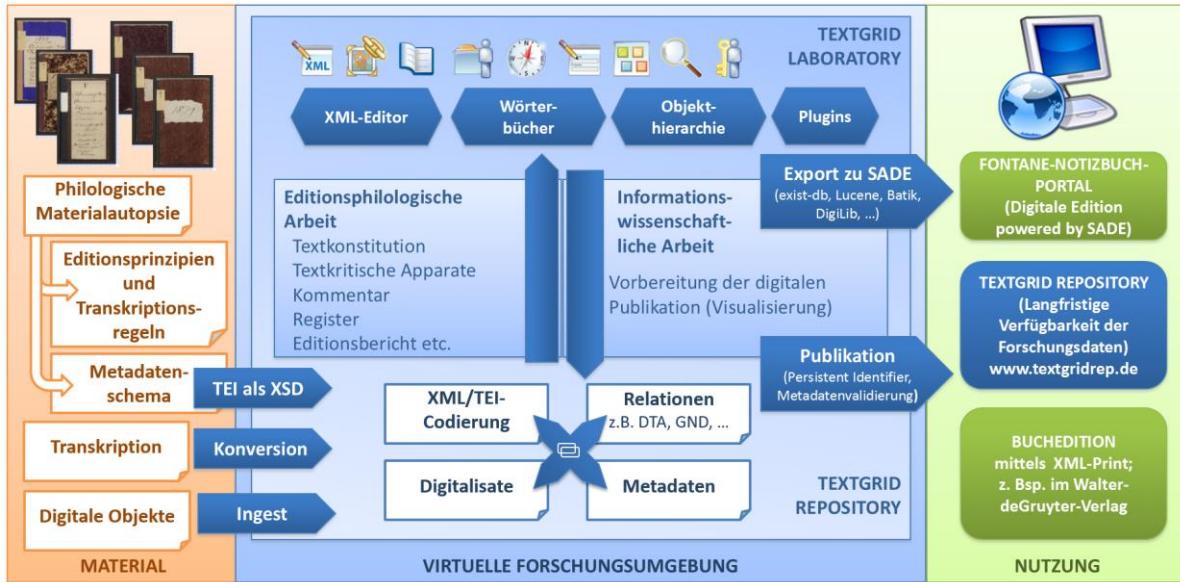


Abbildung: Angepasster und erweiterter TextGrid-Workflow; Grafik erstellt vom TextGrid-Team, ergänzt und bearbeitet für das Fontane-Notizbuch-Projekt von Gabriele Radecke, Martin de la Iglesia und Mathias Göbel

Literaturhinweise:

- Gabriele Radecke: Theodor Fontanes Notizbücher. Überlegungen zu einer überlieferungs-adäquaten Edition. In: Materialität in der Editionswissenschaft. Hrsg. von Martin Schubert. Berlin 2010 (Beihefte zu editio, Bd. 32), S. 95–106.
- Gabriele Radecke: Notizbuch-Editionen. Zum philologischen Konzept der Genetisch-kritischen und kommentierten Hybrid-Ausgabe von Theodor Fontanes Notizbüchern. In: editio 27 (2013). [Im Druck; erscheint im Januar 2014.]
- Gabriele Radecke, Mathias Göbel und Sibylle Söring: Theodor Fontanes Notizbücher. Genetisch-kritische und kommentierte Hybrid-Edition erstellt mit der Virtuellen Forschungsumgebung TextGrid. In: Evolution der Informationsinfrastruktur: Forschung & Entwicklung als Kooperation von Bibliothek und Fachwissenschaft. Hrsg. von Heike Neuroth u. a. Göttingen. [Im Druck; erscheint im Dezember 2013.]

Vortrag:

- Genetisch-kritische und kommentierte Hybrid-Edition von Theodor Fontanes Notizbüchern. Hrsg. von Gabriele Radecke. Gehalten von den Projektmitarbeitern Martin de la Iglesia (SUB Göttingen) und Judith Michaelis (Fontane-Arbeitsstelle Universität Göttingen) auf dem TextGrid-Summit am 15. Mai 2012. In:

<http://www.textgrid.de/fileadmin/praesentationen/tg-summit-2012/praesentation-fontane.pdf>

Alle weiteren Vorträge und Exposés sind auf der Projektwebsite <http://www.uni-goettingen.de/de/303691.html> zusammengestellt.

Kontakt:

Gabriele.Radecke@phil.uni-goettingen.de

martin.de-la-Iglesia@sub.uni-goettingen.de

und

goebel@sub.uni-goettingen.de

DH@Passau – Präsentation des Veranstalters

Die Universität Passau als Veranstalterin möchte mit diesem Poster ihre Forschungsaktivitäten in den Digital Humanities, die nicht erst mit der Einrichtung der beiden Lehrstühle Digital Humanities (April 2013) und Digital Libraries (voraussichtlich April 2014) begannen, präsentieren. Hervorgehoben werden hierbei unter anderem die Kooperationsprojekte „Teuthonista Goes Unicode“, „Die ältesten Ortsnamen im bayerisch-tschechischen Grenzraum (Freyung-Grafenau/Prachatitz)“ und „Diachronic Markup“ sowie weitere Vorhaben aus den Bereichen Geschichte, Kunstgeschichte, Literatur, Sprache und Architektur. Außerdem soll der zum Wintersemester 2013/14 neu eingerichtete Studienschwerpunkt Digital Humanities und dessen Integration in das bestehende, interdisziplinäre Studienprogramm der Universität Passau vorgestellt werden.

DHd 2014, 25.-28.März 2014, Universität Passau

Abstract zum Vortrag:

Was ich nicht weiß, ... macht mich heiß: Zum Mehrwert der Anwendung informatischer Methoden bei der Analyse von Textkorpora am Beispiel des Projektes „Biblia Hebraica transcripta“

Christian Riepl, IT-Gruppe Geisteswissenschaften, LMU München

Der Vortrag betrachtet das Mitte der 1980er Jahre von Wolfgang Richter an der LMU München initiierte Projekt „Biblia Hebraica transcripta“ (BHT) im Licht der aufkommenden „Digital Humanities“ und versucht unter den Aspekten a) Interdisziplinarität, b) Theoriebildung und Methodik sowie c) Gegenstand und Kollaboration, den Mehrwert der Anwendung informatischer Methoden in einem geisteswissenschaftlichen Langzeitprojekt herauszustellen.

a) Interdisziplinarität:

Das Projekt ist von Beginn an wesentlich geprägt durch eine enge und langjährige Kooperation einer geisteswissenschaftlichen mit informatischen Disziplinen. Über die einzelnen Projektförderphasen zwischen 1986 bis 1998 hinaus waren die entwickelten Systeme zum Teil bis zum Jahr 2010 im Einsatz. Der Datenbestand ist weiterhin system-, plattform- und programmunabhängig der Forschung zugänglich. Er umfasst Texte, systematisiertes Grammatikwissen und Ergebnisse der sprachwissenschaftlichen Analyse.

Die Zusammenarbeit führte auf beiden Seiten zu Aus- und Rückwirkungen. Das Forschungsinteresse seitens der Informatik war zunächst begründet in den großen Mengen an Text- und Metadaten sowie den darauf anzuwendenden komplexen Regeln der Grammatik, die sich in den Schwerpunktbereichen „Logikprogrammierung“, „Deduktive Datenbanken“ und „Expertensysteme“ z.B. mit Auswertungsstrategien von Logikprogrammen und der Analysemethode befassste. Im Rahmen der Schwerpunkte „Informationretrieval“ und „Netzzugang zu multimedialen Informationssystemen“ war zunächst die Suche in Feature-Baum-Datenbanken, später der webbasierte Zugang zu Datenbanken interessant.

Auf der anderen Seite war in der althebraistischen Sprachwissenschaft die Übernahme der informatischen, streng formalisierten Denk- und Herangehensweise an einen Gegenstand grundlegend. Die Auswirkungen sind sichtbar z.B. bei der Kodierung der Transkriptionszeichen, der Wahl des Zeicheninventars zur Kodierung linguistischer Annotationen, der logischen und eindeutigen Strukturierung aller Daten, dem Entwurf von Datenbanken und der für die Entwicklung von Analyseprogrammen erforderlichen Formalisierung von Grammatikwissen in Regeln. Erkennbar ist der Einfluss der Informatik weiter an allen Arbeitsschritten der sprachwissenschaftlichen Analyse auf allen methodischen Ebenen, von der Anwendung der Analyseprogramme in einem halbautomatischen Verfahren auf Wort-

und Wortfügungsebene ausgehend bis hin zur datenbankgestützten manuellen Analyse der Satzebene. Schließlich umfasst der Einfluss der Informatik auch das Gebiet der Publikation der Daten, zunächst durch die automatische Konvertierung der Daten zur Verwendung des Drucksatzprogrammes TeX, sodann durch die Entwicklung und den Einsatz von webbasierten Datenbanken auf der Grundlage einer Server-Client-Architektur.

Das Projekt BHt gilt damit in vielerlei Hinsicht und für viele Projekte in den Geisteswissenschaften der LMU als modellhaft für den Einsatz informatischer Methoden und Technologien in geisteswissenschaftlichen Disziplinen.

b) Theoriebildung und Methodik:

Der althebräische Text des Alten Testamentes sollte zunächst transkribiert und in den Computer eingegeben werden, um anschließend durch Computerprogramme grammatisch analysiert zu werden. Die theoretische und methodische Grundlegung für eine orthographiebezogene, morphologisch-syntaktische Transkription war durch W. RICHTER, Transliteration und Transkription: ATSAT 19 (1983) geschaffen. Die Wahl der Datenstrukturen mit Referenzsystem, Segmentierung und Tokenisierung geschah in Abstimmung mit der Informatik, um zum einen größtmögliche Kompatibilität zwischen den Projektpartnern zu erreichen und zum anderen die Daten für eine automatische Analyse vorzubereiten. Basierend auf einem ebenenspezifischen Grammatikmodell, das durch W. RICHTER, Grundlagen einer althebräischen Grammatik, Band 1-3: ATSAT 8 (1978), ATSAT 10 (1979), ATSAT 13 (1980) systematisch begründet war, wurde das dort an einem Ausschnitt des Alten Testamentes gewonnene Grammatikwissen in formale Regeln überführt, die wiederum als Computerprogramme formuliert wurden. Dabei entspricht RICHTER (1978) dem Programm SALOMO zur Analyse der Morphologie, und RICHTER (1979) dem Programm AMOS zur Analyse der Morphosyntax. Beide Analyseprogramme arbeiten kontextunabhängig, streng auf die jeweilige methodische Ebene (Wort bzw. Wortfügung) bezogen und ohne Lexikon. Methodisch setzt die Analyse auf der Wortebene an und schreitet dann über die Wortfügungsebene zu den höheren Ebenen (Satz und Satzfügung) fort. Ein grammatisches, ebenenbezogenes Lexikon entsteht bei der sukzessiv wortweise vorgehenden Analyse der Texte.

Beachtenswert sind nun gerade die unerwarteten Mehrdeutigkeiten, die sich aus der automatischen Analyse auf jeder Beschreibungsebene ergeben. Sie zeigen alle Deutungen auf, die entweder tatsächlich zutreffen können, oder die der Präzisierung, Differenzierung bzw. Einführung weiterer grammatischer und/oder semantischer Regeln bedürfen. Der Analyseprozess legt damit den Erkenntnisweg bei der Sprachbeschreibung offen und zwingt den Experten zu einer Entscheidung, die er auf Grund seiner Kenntnis der jeweils höheren

Beschreibungsebenen reflektiert vornimmt und über ein den Analyseprogrammen nachgeschaltetes Dialogsystem eingibt.

Weiter ist beachtenswert, dass alle in SALOMO und AMOS implementierten Grammatikregeln im Expertendialog konsequent auf den gesamten Textkorpus angewendet worden sind. Die Methodik sah vor, alle im Expertendialog entstandenen Analyseergebnisse aufzuheben und eine Neuberechnung durch die Analyseprogramme nur im Bedarfsfall, z.B. für den Test einer grammatischen Regel oder für ein Experiment durchzuführen. Alle Analyseergebnisse befinden sich neben dem transkribierten Textkorpus in einer relationalen Datenbank, deren Schema versucht, das ebenenspezifische Grammatikmodell in je einer Relation für die Wort-, Wortfügbungs-, Satz- und Satzfübungsebene abzubilden.

Als Ergebnis allein dieser Projektphase liegen vor: Ein vollständig transkribierter, morphologisch und morphosyntaktisch analysierter bzw. annotierter Textkorpus des gesamten hebräischen Alten Testaments. Aus dem Datenmaterial ist eine vollständige Theorie der Morphologie (Bauformen, Wortarten, Kernseme) und der Morphosyntax (Wortfügbungsarten, rekursive Wortverbindungen) herleitbar. Zugleich bildet das Datenmaterial ein grammatisches Lexikon. Konzeptionell sind die Analyseprogramme modifizierbar und somit der Datenbestand unter geänderten Voraussetzungen neu berechenbar. Daneben bietet die Datenbank Möglichkeiten, den Datenbestand unter bestimmten Bedingungen abzufragen, Regeln zu verifizieren bzw. zu falsifizieren, Experimente durchzuführen und neues Wissen zu deduzieren.

Die Anwendung informatischer Methoden führt erfahrungsgemäß zu Überraschungen und Nebeneffekten, die wiederum ein neues Licht auf den Gegenstand werfen und Erklärungen verlangen.

c) Gegenstand und Kollaboration:

Während seiner knapp 30-jährigen Dauer hat das Projekt BHt verschiedene Phasen mit jeweiligen Schwerpunkten durchlaufen. Beispielsweise wurden nach Abschluss der Analysearbeiten die Programme SALOMO und AMOS nicht weiter gewartet. Der Versuch einer automatischen Analyse der Satzebene wurde bislang nicht weiter verfolgt. Ebenso steht das Informationretrievalssystem für Feature-Baum-Strukturen nicht mehr zur Verfügung. Andererseits wurde der gesamte Datenbestand in einer relationalen Datenbank mehrmals reorganisiert. Die Datenbank bhtdb2 konnte schon in einer Frühphase der Entwicklung von Webtechnologien über das webbasierte Informationssystem MultiBHT in erster Linie für Recherchen genutzt werden. Eine soziale Komponente war insofern enthalten, als einfache Benutzerkommentare zu Tokens angebracht werden konnten. Die veraltete und seit 2010 nicht mehr lauffähige Systemtechnologie der Datenbank- und Webschnittstelle wird zurzeit einem umfangreichen Reengineering, das auch kollaborative Aspekte berücksichtigt,

unterzogen. Das Projekt BHt befindet sich mit der nahezu abgeschlossenen rechnergestützt-manuell durchgeführten syntaktischen und semantischen Analyse der Satzebene am Übergang zur dritten Forschergeneration. Ein weiteres Projekt zum althebräischen/semitischen Onomastikon kann ab Frühjahr 2014 auf dem Datenbestand aufbauen.

Als Essenz eines genuinen Langzeit-DH-Projektes lässt sich beobachten:

Den primären Kern digitaler Projekte bildet der gesamte in logisch eindeutig strukturiertem Format vorliegende Datenbestand. Sekundär, weil von der jeweiligen Fragestellung abhängig und daher austauschbar, sind alle digitalen Analyse-, Präsentations- und Recherchewerkzeuge, wobei sich der für diese erforderliche Entwicklungsaufwand durch die rasch fortschreitenden Technologien reduziert, sich Standards herausbilden und digitale Werkzeuge auf andere Problemstellungen übertragbar werden.

Neben der Veröffentlichung von Forschungsergebnissen in Buchform, mit denen Projekte in der Regel ihren Abschluss finden, stehen die Primärdaten eines digitalen Projektes über die eigentliche Projektlaufzeit hinaus weiter zur Nutzung zur Verfügung. Digitale Projekte sind somit nie abgeschlossen und entfalten eine ihnen eigene Dynamik.

Auf dem Weg der Entwicklung hin zu einer kollaborativen Forschungsumgebung stehen der Gegenstand und die Analyseergebnisse objektiviert, transpersonalisiert und universal transformierbar, damit unabhängig von Projekt, Plattform und Forschern im Mittelpunkt. Die Daten können von verschiedenen Forschern verändert und z.B. auch um konkurrierende Meinungen ergänzt, aus anderen Perspektiven, mit anderen Theorien und Methoden, aber auch durch andere Disziplinen untersucht, oder mit digitalen Daten anderer Disziplinen verknüpft werden. Zudem bietet ein Transfer von Projektdaten in Formate und Infrastrukturen von CLARIN und DARIAH hervorragende Möglichkeiten für den Datenaustausch und fördert damit nationale und internationale Kooperation. Der digitale Forschungsgegenstand wird kollaborativ, interdisziplinär und multidimensional erfasst und aus pluraler Sicht betrachtet werden können.

DH? Gibt's doch gar nicht?!

Integration oder Desintegration der Digital Humanities in Deutschland.

Die Digital Humanities (DH, ehemals „Humanities Computing“) als traditionsreiches Forschungsfeld und als Spezialdisziplin haben in ihrer Geschichte schon viele Entwicklungszyklen und Trendwenden erfahren. Wir erleben im allgemeinen Hype-Zyklus derzeit einen noch nie gesehenen Hochstand des Interesses an den Digital Humanities. Dedizierte Förderprogramme zeigen, dass die „digitalen Geisteswissenschaften“ inzwischen auch in den eher politischen Bereichen der Wissenschaft ernst genommen werden. Die schier unüberschaubare Flut von Tagungen wiederum belegt, dass viele neue Akteure das Feld für sich entdeckt haben oder ihre Aktivitäten und Interessen nun unter dieses Label stellen.

Nachdem innerhalb der DH im engeren Sinne schon vor einigen Jahren mit dem Begriff des „big tent“ ein integrativer und inklusiver Kurs eingeschlagen worden ist, scheinen sich nun auch viele Bereiche, die am Rande des Feldes liegen, konzeptionell und begrifflich anzuschließen. Zu den ermutigenden Zeichen gehört hier, dass inzwischen auch speziellere Bereiche wie die Computerlinguistik oder die Archäoinformatik, die lange ihre Eigenständigkeit gepflegt haben, die Nähe zu und den Austausch mit den DH suchen und z.B. verstärkt auf den zentralen DH-Konferenzen auftreten. Auf der anderen Seite gewinnen digitale Verfahren in traditionellen geisteswissenschaftlichen Fächern an Reife und an Bedeutung. Ihre Anwender sehen sich dabei zunehmend selbst in der Nähe oder als Teil der DH. Hier ist z.B. an die digitale Kunstgeschichte oder an die digitale Geschichtswissenschaft zu denken, die zwar Strukturen innerhalb ihrer Fächer aufbauen, den Bezug zu den DH insgesamt aber durchaus sehen. Eine integrative Kraft geht zusätzlich von den großen europäischen Infrastrukturprojekten aus, die zwar von den DH im engeren Sinne angetrieben werden, ausdrücklich aber auf *alle* Geisteswissenschaften zielen, soweit sie mit digitalen Daten arbeiten oder digitale Verfahren einsetzen.

Wie hier bereits implizit angewandt, können die DH als breites Feld beschrieben werden, auf dem es einerseits einen Kernbereich der „DH im engeren Sinne“ bzw. die „DH als eigenes Fach“ gibt und andererseits das Gebiet der „DH im weiteren Sinne“ und die „DH als Methode und Praxis“ in den geisteswissenschaftlichen Fächern. Um das Feld zu kartieren besteht zudem ein Drei-Sphären-Modell, das DH (1.) in einem gleichnamigen Kern, (2.) in „transformierten Fächern“ (Computerphilologie, Digital History) und (3.) in den bestehenden traditionellen Fächern verortet. Mit dem bewusst auf Trennschärfe verzichtenden Modell wird schließlich auch beschrieben, dass die DH ein Bereich „zwischen“ den Geisteswissenschaften und der Informatik bzw. Informationswissenschaft ist und dass es hier um fachlich-konzeptionelle „Bewegungen“ geht, die entweder von den Inhalten zur Technik oder umgekehrt verlaufen.

Mit diesen Ansätzen lassen sich die integrativen Tendenzen der DH auch konzeptionell gut beschreiben. Geht man zu den empirischen Befunden über, dann zeigen sich in letzter Zeit allerdings auch starke desintegrative Momente. Diese nehmen ihren Ausgangspunkt in der Außen- und Selbstwahrnehmung verschiedener klar abgegrenzter Bereiche und Akteursgruppen. Während der Status der DH als eigenständiger Disziplin vor allem empirisch und soziologisch unbestreitbar erscheint, werden eHumanities und Digital Humanities überraschender Weise von verschiedenen

Akteuren zuweilen als entweder nicht-existent oder als überflüssig markiert. Dieses Phänomen begegnet in der Forschungsförderung, wird dann aber vor allem von Akteuren aus der Informatik und aus den traditionellen Geisteswissenschaften getragen. In der Informatik herrscht oft immer noch ein Unverständnis gegenüber der Spezifität und damit auch spezifischen Komplexität geisteswissenschaftlicher Problemstellungen sowie eine gewisse Ignoranz gegenüber den Lösungsansätzen, Theorien, Methoden, Praktiken und Standards, die in den Digital Humanities in den letzten Jahrzehnten entwickelt worden sind. Die unterschwellige Botschaft scheint hier oft zu lauten „Was Informatiker können, können nur Informatiker“. DH wird hier auf eine angewandte Informatik reduziert, die man getrost den Informatikern überlassen sollte. Auf der anderen Seite, bei den geisteswissenschaftlichen Fächern, wird DH als Werkzeugkasten verstanden, der in der Forschung allmählich Eingang findet. Nach der erfolgten Transformation der Fächer in selbstverständlich digital arbeitende Forschungsbereiche wären die DH dann obsolet und könnten wieder verschwinden. Von beiden Seiten wird damit übersehen, dass die DH zwar von Fragestellungen aus den Geisteswissenschaften angetrieben werden und ihren Schwerpunkt in der Entwicklung von Lösungen für die Forschung haben, sich darin aber weder erschöpfen, noch ihren eigentlichen Kern haben. Dieser liegt vielmehr in einem allgemeinen methodologischen Programm, das aus den digital bedingten methodischen Wandlungen der Forschung heraus auf eine neue Epistemologie der Geisteswissenschaften zielt. Die Digital Humanities können in diesem Sinne auch als Interdisziplin bzw. Metadisziplin beschrieben werden.

Für eine produktive Weiterentwicklung der Digital Humanities sind Fragen der Definition und der Bestimmung ihres Gegenstandes von grundlegender Bedeutung. Genauso wichtig ist aber die Formierung der DH als wissenschaftlicher Community. Dabei sind theoretisch zunächst die drei bereits angedeuteten Zuschnitte mit ihren jeweils anderen personellen Zuordnungen und daraus folgenden politischen Agenden möglich: (1.) die DH als eigene Disziplin, (2.) die DH als „transformierte“ Disziplinen und (3.) die DH als Teilbereich der bestehenden Disziplinen. Eine Verengung auf nur eine der drei Interpretationen würde längst der Wirklichkeit der vielfältigen DH-Landschaft widersprechen, ist insofern müßig und würde auch von den Zielstellungen her kontraproduktiv sein. Denn so wie es auf der einen Seite wichtig ist, digitale Methoden und Werkzeuge in die Geisteswissenschaften hineinzutragen, um ihre Transformation zu fördern und zu begleiten, so ist es auf der anderen Seite notwendig, die spezielle Theorie, Methodologie und technische Kompetenz der DH auszubauen, die jeweils über den Horizont der einzelnen Disziplinen hinausgehen muss.

Bei der Bestimmung der DH als Forschungsfeld, als Fachgemeinschaft und damit auch als wissenschaftspolitischer Akteur muss es darum gehen, diese Breite des Feldes zu akzeptieren, die verschiedenen Interpretationen in einem weiten Verständnis von DH zusammenzuführen und trotzdem in einer gemeinsamen Idee zu vertreten. Bei den unterschiedlichen Positionierungen der einzelnen Akteure und Gruppen darf es nicht nur um Fragen der Selbstbehauptung, Abgrenzung und Teilhabe an Fördermitteln gehen. Vielmehr sollte eine gemeinsame Diskussion dazu führen, das Selbstverständnis und Verständnis der DH als Forschungsfeld und als Community zu schärfen und das richtige Maß an Integration nach innen und Trennschärfe nach außen zu etablieren. Vom Ausgang dieser Debatte hängt dann auch ab, ob die Digital Humanities in Deutschland weiter als Teil einer globalen Community auftreten oder ob ein nationaler Sonderweg beschritten wird, der zu einer Abkopplung von den Entwicklungen im Rest der Welt führen könnte.

Der Beitrag berichtet von verschiedenen impliziten und expliziten Positionierungen, Definitionen und Abgrenzungen der Digital Humanities der letzten Zeit. Er macht Vorschläge für eine integrative Kartierung und Beschreibung des Feldes als Grundlage für ein gemeinsames Verständnis der verschiedenen Akteure.

Informationsressourcen

Bekanntmachung des Bundesministeriums für Bildung und Forschung von Richtlinien zur Förderung von Forschungs- und Entwicklungsvorhaben aus dem Bereich der eHumanities. 8. Januar 2013.
<http://www.bmbf.de/foerderungen/21126.php>

Defining Digital Humanities - A Reader. Hrsg. von Melissa Terras, Julianne Nyhan and Edward Vanhoutte. Ashgate 2013.

Sahle, Patrick: Computational Philology? DHd-Blog, 10. Dezember 2013. <<http://dhd-blog.org/?p=2719>>

Sahle, Patrick: DH Studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities. DARIAH-DE Working Papers Nr. 1. Göttingen: GOEDOC 2013.
<http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2013-1.pdf>

Sorting the Digital Humanities Out. Workshop, Universität Umeå, 5.-6. Dezember 2013.
<http://www.humlab.umu.se/sortingdhout/>

Thaller, Manfred: Controversies around the Digital Humanities: An Agenda. In: Controversies around the Digital Humanities. Hg. von Manfred Thaller. HSR Special Issue 37/3 (2012). S. 7-22.

„Erinnerungskultur(en) im World Wide Web – digitale Werkzeuge zur Suche, Dokumentation und Analyse zeithistorischer Narrative im Internet“

Verändert das World Wide Web unseren Umgang mit der Geschichte? Welche Akteure deuten unsere Vergangenheit im Internet mit welchen (Hinter-) Gedanken und Gefühlen? Welche Geschichte(n) erzählen sie wem auf welchen Plattformen?

Wie kann sich die Geschichtswissenschaft mit der Informatik vernetzen, um die digitale Erinnerungskultur zu erforschen?

Als das TV-Event „Unsere Mütter, unsere Väter“ im Frühjahr 2013 ganz ungewohnte Facetten zum Schicksal der Deutschen im Zweiten Weltkrieg präsentierte, eröffnete es eine hitzige Debatte zwischen Wissenschaftlern, Medienschaffenden und Fernsehpublikum über die „korrekte“ Deutung der Vergangenheit. Der Streit überwand die traditionellen Grenzen zwischen Fernsehsessel und Expertendiskurs, er provozierte eine Kaskade von Wortmeldungen in den neuen Medien: Die virtuellen Partizipationsformen reichten von Meinungsäußerungen auf der filmeigenen Facebook-Seite, intensiven Diskussionen auf Twitter unter dem Hashtag „#umuv“, Kontroversen in Blogs, Stellungnahmen im ZDF-Online-Forum „Wie hätten Sie gehandelt? Reden Sie mit!“ über Berichte in Online-Nachrichtenmedien (etwa „Spiegel-Online“) bis hin zu in Online-Mediatheken archivierten TV-Talkshows mit Zeitzeugen.

Die Spezifika des Internets erlauben individuelle Deutungsmuster der Geschichte und deren grenzenlose digitale Distribution. Opfer können zu Tätern umgedeutet, aus geschichtsklitternder Fiktion schnell der Anschein faktischer Erinnerung gemacht werden. Das muss unter historischen und gesellschaftlichen Aspekten unser Interesse erregen.

Zur Aufklärung dieses und weiterer Kardinalprobleme gesellschaftlicher Erinnerung im Zeitalter des World Wide Web schließen sich Magdeburger Wissenschaftler der Fakultät für Humanwissenschaften und der Fakultät für Informatik zu einem gemeinsamen Forschungsvorhaben zusammen.

Für Gesellschaften, die ihre Geschichte aus den Bedürfnissen der Gegenwart (re-) konstruieren, bieten sich im digitalen Raum Möglichkeiten und Risiken gleichermaßen, mediale Verhandlung kollektiver Erinnerung neu zu bestimmen. Doch wir wissen bislang nichts über die Prozesse, die mit der Produktion und der Konsumption dieser memorialen Botschaften verknüpft sind. Es fehlen Untersuchungen zu den Wegen und Karrieren, die solche Geschichte(n) in den digitalen und interaktiven Medien nehmen.

Folgende Epizentren der digitalen Revolution werden wir untersuchen:

- Die Akteure: Wer schreibt die Geschichte, wenn im digitalen Raum keine Grenzen mehr zwischen Experten der gesellschaftlichen Erinnerung (Historiker, Archivare, Museologen etc.), Zeitzeugen und Laien gezogen sind? Wer erringt im Internet die Deutungshoheit über die Vergangenheit? Werden Zeitzeugen gewissermaßen unsterblich, weil digitale Datenbanken im Internet ihr Zeugnis für alle Zeit konservieren?
- Die Narrative: Welche Geschichte(n) werden im Internet wie konstruiert und verbreitet? Welche Deutungen der Zeitgeschichte werden im digitalen Raum

kommuniziert? Auf welche Vergangenheitsnarrative können sich die Akteure des Social Web einigen?

- Die Formen: Wie beeinflussen hypertextuelle Darstellungsmöglichkeiten die Verhandlung von Geschichte? Welche Rolle spielen Bilder und Audiovisionen bei der medialen Verhandlung der Vergangenheit, wenn deren Einsatz technisch kaum noch Grenzen unterliegt? Wie verändern sich die Geschichten über die Vergangenheit, wenn zwischen authentischem Bild und digital konstruierter Fälschung nur ein Mausklick liegt?

Die Aushandlung zeitgeschichtlicher Themen sichtbar zu machen, systematisch zu dokumentieren und zu analysieren, ist die Zielstellung des Magdeburger Digital Humanities-Projekts.

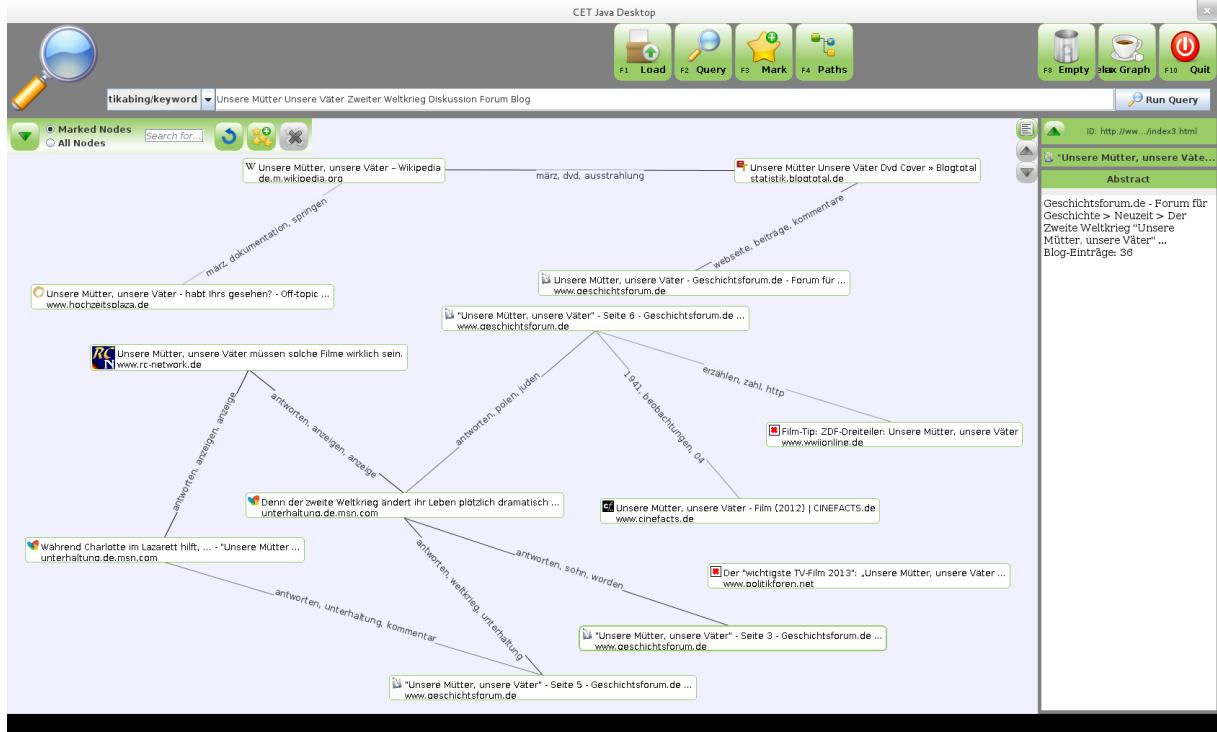
Der Informatik fällt dabei die Aufgabe zu, Methoden und Werkzeuge zu entwickeln, um die extrem umfangreichen und dynamischen, stark vernetzten und zugleich heterogenen digitalen Informationsbestände im Internet zu erheben und sie der Erforschung durch die Geschichtswissenschaften zugänglich zu machen. Dies umfasst eine Unterstützung beim Suchen, Dokumentieren, Archivieren, Bewerten und Analysieren von digitalen Inhalten. Für die damit einhergehenden informationstechnologischen Herausforderungen müssen geeignete Methoden aus den Bereichen Information Retrieval, Machine Learning und Computerlinguistik kombiniert und angepasst werden. Besondere Aufmerksamkeit widmen wir den folgenden Fragestellungen:

- Wie können Aushandlungsprozesse von Erinnerung informationstheoretisch modelliert und mit Hilfe von syntaktischen und semantischen Analysen der Nutzerbeiträge erkannt und untersucht werden?
- Wie können die daran beteiligten Akteure bzw. Nutzer und Nutzergruppen charakterisiert, im Web mit Verfahren aus dem Bereich des Maschinellen Lernens identifiziert, (semi-) automatisiert klassifiziert und plattformübergreifend verfolgt oder auffindbar gemacht werden?
- Können Indexierungs- und Rankingverfahren (z.B. das Vektorraummodell) angepasst werden, um ein effizientes Suchen nach digitalen Erinnerungsinhalten und spezifischen Diskursverläufen zu ermöglichen?
- Welche Visualisierungsmethoden erlauben es, sowohl das gesamte Geflecht aus Akteuren und Narrativen in den unterschiedlichen multimedialen Formen zu veranschaulichen, zeitliche Veränderungen aufzuzeigen und gleichzeitig einzelne Beziehungen zwischen Nutzern, Nutzergruppen und Inhalten greifbar zu machen?

Ziel des gemeinsamen Forschungsprojektes ist die Entwicklung einer Software, die repräsentative geisteswissenschaftliche Untersuchungen von digitalen Inhalten erst ermöglichen.

Im Rahmen dieses Posters wird ein vorführbarer Prototyp vorgestellt, der es erlaubt, die Ergebnisse einer Websuche als Netzwerkstruktur darzustellen. Im Gegensatz zur klassischen Websuche werden hier systematisch Beziehungen (visualisiert als Kanten) zwischen den einzelnen Suchergebnissen (Knoten) gebildet. Diese Relationen lassen sich automatisch über austauschbare semantische oder

statistische Ähnlichkeitsmerkmale bestimmen. Ziel ist es, die komplexen Beziehungen zwischen Akteuren und Narrativen und deren digitale Fußabdrücke systematisch als Netzwerkstruktur zugänglich zu machen. Damit soll sowohl ein Überblick über die Struktur entstehen als auch einzelne Pfade direkt verfolgbar werden.



Die Abbildung zeigt eine solche Netzwerkstruktur am Beispiel zweier Websuchen zum Spielfilm „Unsere Mütter, unsere Väter“. Anhand eines statistischen Ähnlichkeitsmaßes werden Beiträge aus Foren, Blogs, Wikipedia und verschiedenen Portalen miteinander verknüpft, sodass die thematischen Beziehungen (z.B. Brennpunkte mit vielen Kanten) unter den Ergebnissen sichtbar werden.

Die gemeinsam erarbeiteten Werkzeuge sollen die Erforschung des World Wide Web auf repräsentative, effektive, analytisch-innovative Weise erlauben und damit die Interessen der Geisteswissenschaften mit dem Know-How und aktuellen Fragestellungen der Informatik in einem Digital Humanities-Projekt gewinnbringend verbinden.

Imagelab - Digitale Bildwerkzeuge in Forschung und Lehre

(vorwiegend *Themenbereich 2, "Digitale Infrastrukturen für die Geisteswissenschaften"*)

Georg Schelbert (Humboldt-Universität zu Berlin, DE) , ID: 1119

Ziele

Imagelab baut am Institut für Kunst- und Bildgeschichte IKB, gemeinsam mit dem Computer- und Medienservice der Humboldt-Universität zu Berlin und weiteren Projektpartnern eine integrierte Umgebung zum individuellen und gemeinschaftlichen Arbeiten mit Bild- und Forschungsdaten in Lehre, Studium und Forschung.

Insbesondere sind dabei bildspezifische Arbeitsformen – von der Zusammenstellung von Bildcorpora bis hin zur individuellen oder Zusammenarbeit in einer Forschungsgruppe an beliebigen Bildausschnitten und der Onlinepublikation von spezifisch annotierten Bildbeständen – im Fokus.

Bestandteile und Arbeitsfelder

Imagelab besteht aus folgenden Arbeitsumgebungen, die sich bereits im Einsatz oder im Aufbau befinden und lokal oder als Online-Service angeboten werden:

- *imeji* (entwickelt von *imeji* – community, bestehend aus Max-Planck-Digital Library, IKB und anderen, <http://www.imeji.org/>), Medienrepository mit flexibler Metadatenverwaltung.
- *Prometheus* (getragen vom *Prometheus*-Verein e.V., <http://prometheus.uni-koeln.de>), Meta-Bilddatenbank, die die Summe zahlreicher kunsthistorischer, archäologischer, ethnologischer u.a. Bilddatenbanken online erschließt.
- *HyperImage* (Leuphana Universität und KT Hybrid Publishing; <http://hyperimage.ws/de/>); Arbeitsumgebung für den Bilddiskurs, zur Edition und Präsentation von Bildannotationen und –verknüpfungen.
- *Digilib* (Max-Planck-Institute für Wissenschaftsgeschichte Berlin und Kunstgeschichte in Rom und anderen, <http://digilib.berlios.de/>); Online-Graphikserver mit Funktionalitäten zur Bildannotation.
- Das Projekt baut neben diesen bestehenden infrastrukturellen Komponenten auch auf weiteren Projekten auf, die bereits einzelne Elemente verbinden (DFG-Projekt *Meta-Image*).

Imagelab als Summe dieser Infrastrukturen soll die verschiedenen Bereiche des Einsatzes von Digitalbildern bei der wissenschaftlichen Arbeit abdecken: Von der Bereitstellung von Bildern für Lehrveranstaltungen über die Erstellung von fachbezogenen Bildverknüpfungen und -annotationen bis zur Anlage persönlicher und gemeinschaftlicher Bilderpools und Bildpräsentationen. Dies sind im einzelnen folgende Funktionsbereiche:

- Speichern von Bildern (*imeji*)
- Bereitstellen von Bildern im Internet (*imeji*)
- Katalogisieren von Bildern mit Metadaten und Verknüpfung mit Normdaten (*imeji*)
- Integrieren von Bildern in Meta-Bilddatenbanken (*Prometheus*)
- Gemeinsames Verwalten von Bildern (*imeji*, *Prometheus*)
- Annotieren von Bildern (*HyperImage*)
- Verknüpfen von Bildern (*HyperImage*)
- Präsentieren von Bildern und Daten (*Prometheus*, *HyperImage*)

Arbeitsweise

Das Projekt arbeitet an der Schnittstelle zwischen Softwareentwicklung, inhaltlicher Projektarbeit und fachbezogener Wissenschaftspraxis. Es versteht sich daher nicht nur als Infrastrukturprojekt, sondern fragt auch nach dem analytischen Mehrwert digitaler Werkzeuge für die Geisteswissenschaften und ihren kommunikativen Funktionen.

Parallel zur technischen Entwicklungstätigkeit, die bei den einzelnen Infrastrukturen in jeweils eigenen Projekten stattfindet, werden in der zweijährigen Projektphase von „*Imagelab*“ neue Arbeitsszenarien entwickelt, um diese Infrastrukturen zu einem stetigen, verschiedenen Disziplinen offenstehenden Angebot weiterzuentwickeln, das vernetztes und kollaboratives Arbeiten mit Bildern und zugehörigen Daten in Forschung, Lehre und Studium ermöglicht.

Hierzu betätigt sich das Projekt in folgenden Gebieten

- Analyse von vorhandenen Tätigkeitsfeldern im bildwissenschaftlichen Bereich
- Entwickeln von Szenarien für den Einsatz bildbezogener digitaler Infrastrukturen in den Geisteswissenschaften
- Vermitteln der verfügbaren Funktionalitäten an ein geisteswissenschaftliches Publikum
- Austausch zwischen Geisteswissenschaftlern und Informatikern

Der Übertragbarkeit der Ergebnisse auf andere Anwendungsbereiche und Einrichtungen wird dabei besondere Aufmerksamkeit geschenkt.

Förderung

Imagelab wird vom Medienkommission-Förderprogramm 2013-2015 der Humboldt-Universität zu Berlin “Digitale Medien in Lehre und Forschung” unterstützt.

Die Einrichtung eines institutionenübergreifenden Kompetenzzentrums für den Einsatz digitaler Bilder in den Geisteswissenschaften *ImageHumanities* ist in Vorbereitung. Hierzu wurde ein Förderantrag beim BMBF gestellt. Falls dieses Projekt bis zur DHd-Jahrestagung bereits entsprechende Realisierungsschritte durchlaufen hat, werden diese in das Poster einfließen.

30.12.2014

Panel: ICE/AGE - Von der Anwendungsinsel zum digitalen Marktplatz und Hörsaal

Dieses Panel wird von der Arbeitsgemeinschaft Geschichte und EDV e.V. (AGE) in Kooperation mit dem Interdisciplinary Center of E-Humanities in History and Social Sciences des Max Weber-Kollegs der Universität Erfurt organisiert (ICE). Das ICE ist ein Verbund aus Forscherinnen und Forschern der Universitäten Erfurt, Graz, Hamburg, Magdeburg, Leipzig, Trier und Ilmenau, des Fraunhofer Instituts für digitale Medientechnologie Ilmenau und der FH Erfurt, zusammengeschlossen als Forschungsstelle am Max-Weber-Kolleg für kultur- und sozialwissenschaftliche Studien der Universität Erfurt. Das Kernziel des Verbundes besteht in der Entwicklung und Verknüpfung digitaler Analyseverfahren und -werkzeuge, die im Spektrum der geschichts-, sozial und kulturwissenschaftlichen Forschung sowie in der forschungsorientierten Lehre (insbesondere in den „Digital Humanities“) anwendbar sein sollen.

In diesem Sinne thematisiert das Panel in einem ersten Schritt neue Methoden der „Digital Humanities“, welche im Rahmen des ICE zum Einsatz kommen. **Prof. Werner Rieß** (Historisches Seminar - Arbeitsbereich Alte Geschichte, Universität Hamburg) präsentiert das ERIS Projekt (ERIS - Hamburg Information System on Greek and Roman Violence). In dem auf MyCore basierenden Informationssystem sollen alle Gewaltbeschreibungen, die sich in den Werken griechischer und lateinischer Autoren finden, aufgenommen und mit spezifischen semantischen Kriterien versehen werden. Alle Passagen, die interpersonelle Gewalt beschreiben oder erwähnen, sollen einer einfachen wie einer erweiterten Suche zugänglich gemacht werden. Neben den offensichtlichen Merkmalen wie Autor, chronologischer Einordnung von Werk und Inhalt, werden viele weitere Eigenschaften von Gewaltakten erfasst. Diese betreffen unter anderem die Kontexte, Motive, geographischen Verortungen, den sozioökonomischen Status und das Alter der jeweiligen Akteure sowie die Folgen eines Gewaltaktes im weitesten Sinne von unmittelbaren Gegenreaktionen bis hin zu gesetzgeberischen Maßnahmen. Durch eine feine Aufgliederung dieser Merkmale von Gewaltakten wird eine zielgerichtete Suche bei größtmöglicher Benutzerfreundlichkeit ermöglicht.

Prof. Christoph Schäfer (FB III - Alte Geschichte, Universität Trier) stellt das AIDA-Projekt (Adaptiver, Interaktiver, Dynamischer Atlas zur Geschichte - Visuelles Erkunden und interaktives Erleben der Geschichte) vor. Hauptziel des Projektes ist die Entwicklung eines datenbankgenerierten, dynamischen und adaptiven Atlas zur Geschichte Europas und des Mittelmeerraumes für Bildung und Forschung. Dynamische Karten ermöglichen die Visualisierung von räumlichen und zeitlichen Veränderungen von „Objekten“ und „Vorgängen“ und vermitteln damit historische Prozesse und Entwicklungen. Durch das Variieren von Abfragekriterien können historische Zusammenhänge adaptiv auf Übersichts- und Detailkarten so dargestellt werden, dass der Atlas selbst zur Quelle neuer Erkenntnisse wird. Die Interaktivität der Karten ermöglicht den direkten Zugriff auf Datenbanken mit Quellenmaterial und Forschungsergebnissen sowie deren Ergänzung im Zuge individueller Forschungsprojekte.

Prof. Charlotte Schubert (Historisches Seminar - Lehrstuhl Alte Geschichte, Universität Leipzig) wird am Beispiel Textmining zeigen, welche Analysemöglichkeiten sich aus der Kookkurrenzsuche ergeben. Insbesondere aus der Kookkurrenzsuche ergibt sich ein anderer Blick auf das Phänomen der Serendipity: Die Suche nach den seltenen Kookkurrenzen hat gezeigt, dass die ‚seltenen Ereignisse‘ die vielversprechenderen Kandidaten für das Auffinden neuer, interessanter Zusammenhänge sind

im Vergleich zu den statistisch häufigeren. Hieraus lässt sich für den Einsatz der Kookkurrenzsuche als methodisches Prinzip ableiten, in der speziellen Form der explorativen Suche die seltenen Kookkurenzen zu betrachten, um genau solche ungewöhnlichen, seltenen und bisher von der Forschung nicht gesehenen oder für abwegig gehaltenen Zusammenhänge aufzudecken.

Nach dieser Vorstellung neuer Methoden und Anwendungen im Kontext des ICE thematisiert das Panel in einem zweiten Schritt die Frage nach der Vereinbarkeit und Kombinierbarkeit der einzelnen Methoden. In einem ersten Beitrag hierzu stellt **Prof. Klaus P. Jantke** (Fraunhofer Institut für digitale Medientechnologie Ilmenau) das Konzept der Meme Media vor. Ausgangspunkt ist die Erkenntnis, dass es auch im nicht-biologischen Bereich Evolution gibt, Mutationen, Kreuzungen, aber auch das Aussterben eingeschlossen. Meme Media bezeichnet eine Familie von technologischen Ansätzen - Objekte einer Model-View-Controller-Architektur - die geeignet sind, mit Mitteln der Informations- und Kommunikationstechnologien die Evolution von Wissen, das digital repräsentiert ist, zu befördern. Meme Media unterstützen Benutzer, Wissensbausteine derart zu manipulieren, dass die Formulierung bisher nicht ausgedrückter Einsichten und Zusammenhänge geradezu provoziert wird. Die Darstellung wird anhand markanter Beispiele illustriert.

Ergänzt wird diese grundlegende Einführung durch einen Beitrag von **Prof. Wolfgang Spickermann** (Institut für Alte Geschichte und Altertumskunde – Karl Franzens-Universität Graz) und **Dr. Leif Scheuermann** (Max-Weber-Kolleg – Universität Erfurt) über den praktischen Mehrwert der Webble Technologie (WEB-Based Life-like Entities) für die „Digital Humanities“, einer webbasierten Ausprägung der Meme Media Technologie, durch welche unterschiedlichste verteilte Anwendungen und Daten auf einer Oberfläche zusammengebracht, sie dynamisch manipuliert und frei miteinander kombiniert werden können, ohne sie selbst dabei verändern zu müssen. Webbles erlauben Nutzern, vorhandene Wissensressourcen, welche als Medienobjekte gekapselt – „gewrapped“ – sind, weiterzuverarbeiten und zu distribuieren. Benutzer können einzelne Medienobjekte durch direkte Manipulation, wie „drag“, „drop“, „copy“, „paste“, miteinander zu neuen Objekten kombinieren, ohne Programmierkenntnisse zu besitzen. Dies erlaubt die einfache und dynamische Verbindung von Methoden z.B. der qualifizierenden Datenanalyse (Textmining) mit GIS-basierten Visualisierungssystemen oder Netzwerkanalysen.

Das Panel wird abgeschlossen durch einen Beitrag von **Dr. Thomas Grotum** (Fb III - Neuere und Neueste Geschichte, Universität Trier), der sich mit der Vermittlung der zuvor thematisierten Methoden beschäftigt. Vor dem Hintergrund bereits existierender und geplanter Studiengänge im Bereich „Digitale Geisteswissenschaften“ geht es um die Anforderungen und den Praxisbezug einer universitären Ausbildung aus Sicht der Geschichtswissenschaft. Es stellt sich die Frage, wie Absolventen eines entsprechenden Masterstudiengangs in die Lage versetzt werden können, fachliche Problemstellungen mit Hilfe rechnergestützter Verfahren und ggf. unter Zuhilfenahme von digitalen Ressourcen methodisch reflektiert zu lösen.

Bestätigung der Teilnehmer

Hiermit Bestätige ich, dass alle genannten Teilnehmer mir ihre Bereitschaft an der Teilnahme am Panel „ICE/AGE - Von der Anwendungsinsel zum digitalen Marktplatz und Hörsaal“ mündlich zugesagt haben, was auch an der Mitgestaltung des Abstracts erkenntlich wird.

MfG

Dr. Leif Scheuermann

1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd)
Universität Passau, 25.-28.03.2014

Vortragsvorschlag

Für eine computergestützte literarische Gattungsstilistik

Christof Schöch, Steffen Pielström

(Lehrstuhl für Computerphilologie, Universität Würzburg)

Einleitung

Der vorliegende Beitrag plädiert für eine computergestützte literarische Gattungsstilistik, verstanden als eine Forschungsagenda für die Literaturwissenschaften, welche hermeneutische und quantitative Methoden verbindet. Diese Agenda wird im Zusammenhang mit einem in Vorbereitung befindlichen Forschungsprojekt zum gleichen Thema formuliert, das in der romanistischen Literaturwissenschaft angesiedelt ist. Aus diesem Forschungsprojekt werden zwei Zwischenergebnisse berichtet: das Erste betrifft die konzeptuelle Verknüpfung von Gattungstheorie und computergestützter Stilistik; das Zweite betrifft die methodische Erweiterung der Principal Component Analysis (PCA) für literaturwissenschaftliche Fragestellungen.

1. Die Agenda der computergesetzten literarischen Gattungsstilistik

Die übergeordnete Zielsetzung einer computergesetzten literarischen Gattungsstilistik ist es, eine tiefgehende Konvergenz herzustellen zwischen etablierten literaturwissenschaftlichen Fragestellungen einerseits und quantitativen Verfahren der Textanalyse andererseits. Eine solche Konvergenz ist Voraussetzung dafür, dass sich entsprechende Forschungsvorhaben im Kernbereich der Digital Humanities ansiedeln können, in dem computergestützte Geisteswissenschaften und Angewandte Informatik nicht nebeneinander stehen, sondern sich zu einem neuen, dritten Forschungsparadigma verbinden.

Auf eine computergestützte literarische Gattungsstilistik bezogen ergeben sich daraus eine Reihe von Forschungsfragen. Einige von ihnen sind primär literaturwissenschaftlich: Wie kann die Beziehung zwischen Stil und Gattung in einer produktiven Weise konzeptualisiert werden? Wie verhalten sich Gattungsstile, Epochentypen und Autorenstile zueinander? Welche anderen Faktoren spielen für die stilistische Beschreibung literarische Texte eine Rolle? Welche automatisch identifizierbaren sprachlichen Merkmale, auf welchen Ebenen der linguistischen Beschreibung, sind Indikatoren für Gattungen? Andere Fragestellungen

stammen aus dem informatischen Bereich des *Text Mining*: Welche Verfahren der Text-Kategorisierung und des Maschinellen Lernens können eingesetzt und angepasst werden? Wie können für Verfahren wie *Support Vector Machines* die besten *kernels* definiert und geeignete *features* modelliert werden? Aus der Verbindung von literaturwissenschaftlicher und informatischer Fragestellungen ergeben sich aber auch ganz neue Fragen, die dem spezifischen Bereich der Digital Humanities zuzurechnen sind: Welche Besonderheiten natürlichsprachlicher und spezifisch literarischer Daten sind zu berücksichtigen, wenn es darum geht, möglichst generische Strategien zur Trennung von Autoren- Epochen und Gattungssignal zu entwickeln? Wie können computergestützte Verfahren so weiterentwickelt werden, dass sie einerseits auch für literatursprachliche Daten statistisch signifikant, robust und verlässlich sind, dass sie andererseits aber auch aus hermeneutischer Perspektive transparent und interpretierbar, das heißt aus literaturwissenschaftlicher Sicht bedeutungsvoll sind? Und allgemeiner, wie verändert die computergestützte Herangehensweise die Weise, wie wir über literarische Interpretation und algorithmische Analyse sowie ihre wechselseitige Beziehung nachdenken? Die Bearbeitung dieser Forschungsfragen bildet den Kern der Forschungsagenda einer computergestützten literarischen Gattungsstilistik. Zu zwei dieser Teilfragen werden hier Zwischenergebnisse berichtet.

2. Die konzeptuelle Verknüpfung von Gattungstheorie und quantitativer Stilistik

Das erste Zwischenergebnis bezieht sich auf die konzeptuelle Verknüpfung von Gattungstheorie und computergestützter Stilistik. Der Gattungsstilistik geht es um einen induktiven, deskriptiven Blick auf die stilistischen Merkmale literarischer Gattungen und Untergattungen sowie auf deren historische Entwicklung.

In der neueren Gattungstheorie hat sich die Auffassung durchgesetzt, dass literarische Gattungen sich nicht mit einem idealistischen, deduktiven Ansatz systematisieren lassen (Schaeffer 1989). Vielmehr sind sie als historische Konventionen zu verstehen, die komplexe und sich dynamisch entwickelnde „generic facets“ (Kessler et al. 1998) umfassen. Diese beziehen sich zu unterschiedlichen Anteilen auf Themen, Plot und diverse stilistische Merkmale (Hoffmann 2009). Die Kombination mehrerer solcher "facets" definiert eine Gattung oder Untergattung, und Übergangsformen oder diachrone Entwicklungen lassen sich über den Wegfall oder das Hinzutreten einzelner "facets" erfassen.

In ähnlicher Weise wird Stil heute als ein Phänomen aufgefasst, das als „Bündel konkurrierender Merkmale“ auf unterschiedlichen linguistischen Beschreibungsebenen (Phonologie, Morphologie, Lexik/Semantik, Syntax, Plot, etc.) verstanden werden kann

(Sandig 2006, Karlsgren & Cutting 1994).¹ Hier kann die computergestützte Stilistik ansetzen, denn sie ist in der Lage, induktiv und umfassend zahlreiche Merkmale - in ihrer gegenseitigen Abhängigkeit, in ihrer jeweiligen Gewichtung, und unter präziser Berücksichtigung zahlreicher möglicherweise relevanter Faktoren - zu erfassen und für die Klassifikation oder das Clustering von Texten zu nutzen. Damit wird die von Dominique Combe eingeforderte „stylistique des genres“ (Combe 2002) computergestützt realisiert. Abb. 1 fasst das Verhältnis zwischen der Theorie literarischer Gattungen und computergestützter Stilistik / Text Mining zusammen.

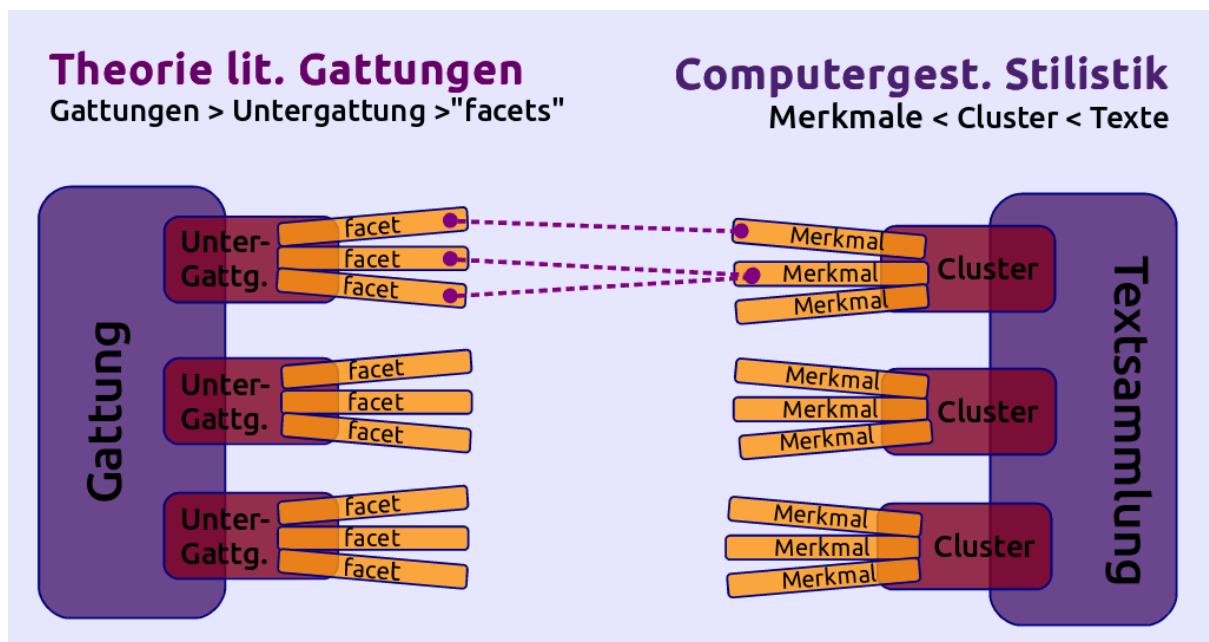


Abb. 1: Literarische Gattungstheorie und computergestützte Stilistik

Durch die vergleichbare Konzeption von Gattungen (mit Facetten) und Stil (mit Merkmalen) können Verbindungen zwischen historisch oder theoretisch gegebenen Untergattungen einerseits und auf der Grundlage stilistischer Ähnlichkeit gruppierten Clustern von Texten andererseits entdeckt werden. Genauer gesagt: es können in ihrer Stärke statistisch charakterisierbare Korrelationen zwischen einzelnen Gattungsfacetten und stilistischen Merkmalen erhoben und eingeordnet werden. Durch die Identifikation von besonders distinktiven Merkmalen und durch Merkmalsgeneralisation können dann auch

¹ Im Bereich der Corpuslinguistik geht die computergestützte Untersuchung der stilistischen Unterschiede von (literarischen) Gattungen und Untergattungen bis in die 1980er-Jahre zurück, mit Pionierarbeiten von Douglas Biber zur Modellierung des Zusammenhangs zwischen funktionalen Gattungsaspekten und stilistischen Merkmalen, die zu synthetischen Dimensionen zusammengefasst werden (Biber 1992) und der Erprobung einer breiten Auswahl von potentiellen "style markers" (Karlsgren & Cutting 1994). Außerdem wurden bspw. die vergleichende Evaluation von token-basierten, syntaktischen und anderen Merkmalen vorgenommen (Stamatatos et al. 2000).

Merkmalsbündel ermittelt werden, die zugleich statistisch signifikant mit einer Facette korelieren und aus literaturwissenschaftlicher Perspektive interpretierbar sind.

3. Die methodische Erweiterung der Principal Component Analysis

Die computergestützte literarische Gattungsstilistik ist Teil einer sich aktuell verstärkenden Tendenz, stilometrische Fragen über die traditionell im Vordergrund stehende Autor-Attribution hinaus zu bearbeiten.² Zahlreiche wohl etablierte Methoden (wie bspw. *Cluster Analysis* oder *Principal Component Analysis*), aber auch neuere informatischen Verfahren aus dem Bereich des *Text Mining* und *Machine Learning* sind für eine so konzipierte Gattungsstilistik anschlussfähig. Wir schlagen hier vor diesem Hintergrund vor, die etablierte Methode der *Principal Component Analysis* (PCA, siehe grundlegend Jackson 2005) auf eine Weise zu erweitern, die ihre Interpretierbarkeit erhöht.

Zur verlässlichen Unterscheidung von Kategorien oder Gruppen innerhalb einer großen Menge von Texten ist die Stilometrie auf die gleichzeitige Betrachtung einer großen Zahl von Merkmalen (bspw. Wörtern) angewiesen. Fasst man jedes Merkmal mehrerer Texte als je eine Dimensionen auf, beruht die stilometrische Erhebung von Ähnlichkeitsrelationen zwischen Texten daher häufig auf einer Dimensionsreduktion. Anders als bspw. Burrows' Delta (Burrows 2002) erlaubt PCA die Reduktion der Dimensionen eines Datensatzes nicht auf nur eine einzige, sondern auf wenige neue und voneinander unabhängige Dimensionen, die sog. *principal components* (PC), die die Varianz in den Daten besonders gut beschreiben. Jede dieser PCs korreliert hierbei mit einer spezifischen Kombination bzw. Gewichtung von Merkmalsfrequenzen. Charakteristisch für die PCA ist, daß bereits die ersten 2-4 PCs oft einen großen Teil der im Datensatz enthaltenen Varianz beschreiben (Abb. 2a). Für eine graphische Exploration der Ähnlichkeit zweier oder mehrerer Gruppen kann es also ausreichen, die ersten PCs als Koordinaten zu verwenden, anstatt in einer großen Zahl von Wortfrequenzen eine Kombination zu suchen, die Unterschiede besonders deutlich macht (Abb. 2b).

² Im Bereich der literarischen Gattungsstilistik liegen mehrere Ansätze vor, welche die diachrone Entwicklung von Gattungen allgemein (Moretti 2005, Jockers 2013) oder spezifische methodische Lösungsversuche betreffen: bspw. Cluster Analyse oder "unmasking"-Prozedur für die Gattungsklassifikation (Allison et al. 2011; Kestemont et al. 2012; Schöch 2013).

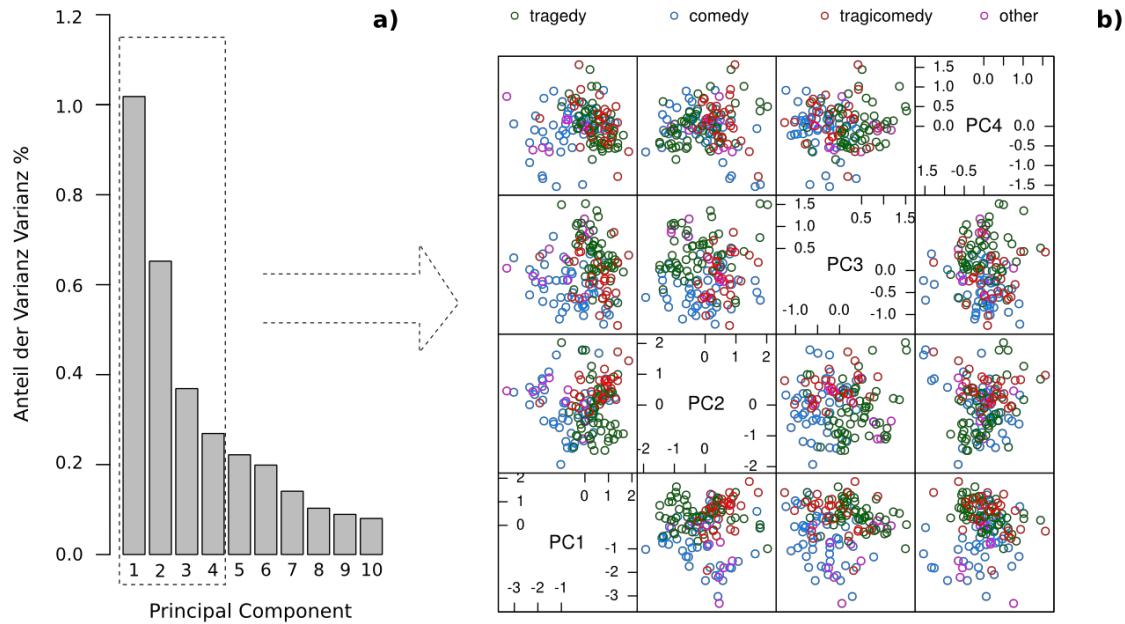


Abb. 2: Die Principal Component Analysis am Beispiel von 141 französischen Dramen (Tragödien, Komödien, Tragikomödien, Andere) aus dem siebzehnten Jahrhundert, basierend auf den relativen Häufigkeiten der 200 am meisten verwendeten Wörter. **a)** Scree Plot mit den Varianzanteilen für PC 1-10.; **b)** Scatterplot Matrix für PC 1-4 mit gattunsspezifischer Farbcodierung.

Die PCA erlaubt allerdings von sich aus keine Aussagen über die Unterschiedlichkeit zweier Kategorien von Datenpunkten, oder über den Einfluß einer bestimmten Gruppierungsvariable, weshalb stilistische Unterschiede zwischen Gattungen mit dieser Methode zwar visualisiert, aber nicht analysiert werden können. Um diese Lücke zu überbrücken, haben wir ein Verfahren entwickelt, mit dem der Einfluß der Zugehörigkeit eines bestimmten Textes zu einer spezifischen Kategorie (bspw. der Gattung) auf die PCs mittels der Varianzanalyse (ANOVA) untersucht werden kann. Hierbei wird die Gattung als unabhängige faktorielle Variable betrachtet, die Werte einer bestimmten PC als abhängige Variable. Das Bestimmtheitsmaß R^2 liefert hierbei eine Vergleichsgröße, anhand derer sich die Einflüsse verschiedener Faktoren (bspw. Gattung, aber auch Autorschaft oder auch Publikationsdatum) auf eine bestimmte PC quantifizieren lassen (Abb 3a).

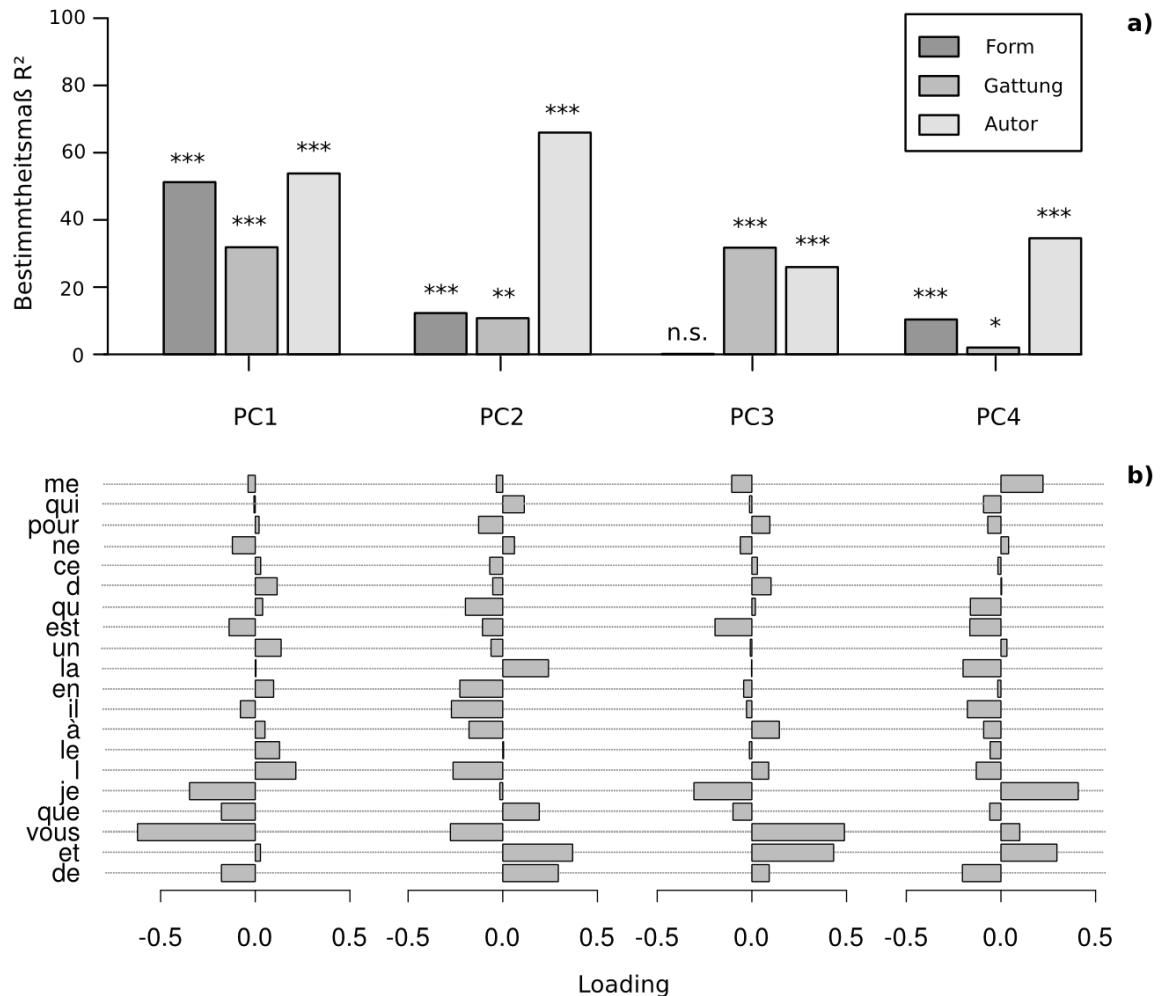


Abb. 3: Von der PCA zur Kategorie. **a)** Ergebnisse der Varianzanalysen (ANOVAs) zur Quantifizierung des Einflusses verschiedener Faktoren auf die PCs 1-4. Untersucht wurden die Faktoren Form (d.h. Vers, Prosa oder Gemischt), Gattung (Tragödie, Komödie, Tragikomödie) und Autor. Dargestellt sind die Bestimmtheitsmaße (R^2) für jeden dieser Faktoren bei jeder PC. Symbole über den Balken repräsentieren das Signifikanzniveau der jeweiligen Beziehung ('n.s.' nicht signifikant; '*' $p < 0.05$; '**' $p < 0.01$; ***' $p < 0.001$). Den stärksten Einfluß hat die literarische Gattung in diesem Datensatz auf PC1 und PC3; **b)** Loading-Werte der 20 häufigsten Worte für PC 1-4. Diese Werte repräsentieren den Einfluß einzelner Wortfrequenzen auf die PCs, und erlauben so in Kombination mit den Resultaten der Varianzanalysen interpretierende Rückschlüsse.

Die Kombination von PCA und ANOVA erlaubt somit die Identifikation von PCs, die durch einen bestimmten Faktor besonders stark geprägt werden, und den Abgleich mit der Gewichtung bestimmter Wortfrequenzen in dieser PC (Abb. 3b). Gemeinsam betrachtet erlaubt dies Rückschlüsse darüber, welche Merkmale und Merkmalsbündel für die Unterschiede in bestimmten Faktoren besonders relevant sind, d.h. auch welche Kombinationen charakteristisch für Gattungsunterschiede sind.

Fazit

Mit der konzeptuellen Verbindung von Gattungstheorie und computergestützter Stilistik einerseits, und der methodischen Erweiterung der PCA zur verbesserten Interpretierbarkeit von PCs in Bezug auf relevante Kategorien andererseits, konnten wichtige Zwischenziele erreicht und Grundlagen für die eingangs beschriebene Forschungsagenda gelegt werden. Und es konnte exemplarisch gezeigt werden, so hoffen wir, wie die hier vertretene Forschungsagenda einer computergestützten literarischen Gattungsstilistik hermeneutische und quantitative Methoden zu einem eigenständigen Ansatz verbindet, der auf die Entwicklung spezifisch erweiterter informatischer Verfahren für ein erweitertes Verständnis der Natur und der Entwicklung literarischer Gattungen abzielt.

Literaturangaben

- Allison, Sarah, Ryan Heuser, Matthew L. Jockers, Franco Moretti, and Michael Witmore (2011). *Quantitative Formalism: An Experiment*. Stanford: Stanford Literary Lab.
- Biber, Douglas (1992). "The multidimensional approach to linguistic analyses of genre variation", in: *Computers in the Humanities*, 26.5-6, 331-347.
- Burrows, John (2002). "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship". *Literary and Linguistic Computing* 17.3, 267-287.
- Combe, Dominique (2002). "La stylistique des genres", in: *Langue française* 135, 33-49.
- Hoffmann, Michael (2009). "Mikro- und makrostilistische Einheiten im Überblick", in: *Rhetorik und Stilistik, Ein internationales Handbuch historischer und systematischer Forschung*, Band 2, hg. von Ulla Fix, Andreas Gardt & Joachim Knape. Band 2, Berlin: de Gruyter, 1529-45.
- Jackson, Edward (2005). *A User's Guide to Principal Components*. New York: Wiley.
- Jockers, Matthew (2013). *Macroanalysis. Digital Methods and Literary History*. Chicago: Univ. of Illinois Press.
- Juola, Patrick (2006). "Authorship Attribution." *Foundations and Trends in Information Retrieval* 1.3, 233-334.
- Karlgren, Jussi & Douglas Cutting (1994). "Recognizing text genres with simple metrics using discriminant analysis". In *Proceedings of COLING '94*, Vol. 2, 1071-1075.
- Kessler, Brett, Geoffrey Numberg, and Hinrich Schütze (1998). "Automatic Detection of Text Genre." In *Proceedings of ACL 1998*, 32-38. doi:10.3115/976909.979622.
- Kestemont, Mike, Kim Luyckx, Walter Daelemans, and Thomas Crombez (2012). "Cross-Genre Authorship Verification Using Unmasking." *English Studies* 93.3, 340–356.
- Moretti, Franco (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.
- Sandig, Barbara (2006). *Textstilistik des Deutschen*. 2. Auflage. Berlin: de Gruyter.
- Schaeffer, Jean-Marie (1989). *Qu'est-ce qu'un genre littéraire?* Paris: Seuil, 1989.
- Schöch, Christof (2013). "Fine-tuning Our Stylometric Tools: Investigating Authorship and Genre in French Classical Theater", *Digital Humanities Conference 2013*, <http://dh2013.unl.edu/abstracts/ab-270.html>.
- Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis (2000). "Automatic Text Categorization in Terms of Genre and Author." *Computational Linguistics* 26/4, 471-497.

Freier Wissenszugang vs. Geistiges Eigentum: Hürden und Lösungsmodelle für Forschung und Lehre im digitalen Raum

Walter Scholger

Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities

Universität Graz

Der Beitrag widmet sich rechtlichen Problemen, mit denen sich GeisteswissenschaftlerInnen im digitalen Raum häufig konfrontiert sehen (wie z.B. Besitzverhältnisse digitaler Reproduktionen, elektronische Zurverfügungstellung von digital(isiert)en Quellen, ...). Dabei wird insbesondere auf gesetzliche Restriktionen und Stolpersteine, aber auch gängige Missverständnisse hingewiesen, die aus der heterogenen internationalen Rechtslage zu Urheberrechten und dem Schutz geistigen Eigentums resultieren. Basierend auf dem Vergleich der unterschiedlichen Rechtsvorschriften im deutschsprachigen Raum und internationaler Beispiele werden *best practice* Beispiele für den Umgang mit diesem Thema definiert.

Der freie Zugang zu wissenschaftlichen Quellen für Lehre und Forschung ist ein zentrales Anliegen aller Disziplinen. Durch die Vorgaben nationaler und internationaler Fördergeber, wie den Auflagen zur öffentlichen und freien Verfügbarkeit von Forschungspublikationen, ist der freie Zugang zu den Ergebnissen wissenschaftlicher Arbeit zu einer unausweichlichen Notwendigkeit geworden. Wie aber steht es mit den Quellen wissenschaftlicher Arbeit? Im Allgemeinen widmet sich geisteswissenschaftliche Forschung Produkten des menschlichen Geistes, so dass die Forschungsobjekte Regelungen zum Schutz des geistigen Eigentums der UrheberInnen unterliegen. Da diese Forschung aber primär an Universitäten, Kulturerbeinstitutionen oder anderen öffentlichen Einrichtungen stattfindet, ist sie üblicherweise nicht kommerziell orientiert, sondern einem öffentlichen Bildungsauftrag geschuldet. Die öffentliche Hand verfügt selten über Mittel zur Lizensierung und Abgeltung von Urheberrechten. Der offene und freie Zugang zu Quellen – insbesondere jenen an Gedächtnisinstitutionen wie Archiven und Bibliotheken – gewinnt zusätzlich an Bedeutung, weil nationale Fördergeber (wie z.B. der österreichische FWF) keine Finanzierung für Lizenzierungen vorsehen.

Andererseits haben ForscherInnen selbst beträchtliches Interesse, ihr eigenes geistiges Eigentum zu schützen – sowohl aus wirtschaftlichen, als auch akademischen Überlegungen. Dieser Konflikt zeigt sich bereits in der Allgemeinen Erklärung der Menschenrechte (Art. 27), in der zunächst „*das Recht, am kulturellen Leben der Gemeinschaft frei teilzunehmen, sich an den Künsten zu erfreuen und am wissenschaftlichen Fortschritt und dessen Errungenschaften teilzuhaben*“ festgeschrieben wird, ehe der nächste Absatz einräumt: „*Jeder hat das Recht auf Schutz der geistigen und materiellen Interessen, die ihm als Urheber von Werken der Wissenschaft, Literatur oder Kunst erwachsen*“.

Eine Reihe von EU Richtlinien (2001/29/EC, 2003/98/EC, 2004/48/EC) zeigen ein klares politisches Bekenntnis für den freien Zugang zu Wissen und den freien Gebrauch von Lehr- und Forschungsmaterialien. Dieses zeigte sich auch in den Förderrichtlinien des 7. Rahmenprogramms der EU, der *EU Digital Agenda for Europe* und Veröffentlichungen der UNESCO (z.B. *Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace* und *Charter on the Preservation of Digital Heritage*, jeweils 2003).

Die nationale Implementierung dieser Agenda in vielen Mitgliedsstaaten der EU ist jedoch alles andere als zufriedenstellend und stellt ForscherInnen (bzw. generell NutzerInnen) vor eine Reihe von Problemen.

Während *Common Law* Rechtssysteme (wie z.B. die USA und Großbritannien mit den Konzepten *Fair Use* und *Fair Dealing*) das Recht der Gesellschaft auf den Zugang zu und den Gebrauch von Veröffentlichungen für Zwecke der Bildung höher bewerten, hat in den *Civil Law* Rechtssystemen Kontinentaleuropas der Schutz der Rechte der UrheberInnen mehr Gewicht. Daher erfordert die Verwendung, Verbreitung und vor allem die elektronische Zurverfügungstellung solcher Ressourcen im Bildungsbereich klare, im Gesetz definierte, Ausnahmen. Während solche Privilegien für analoge Ressourcen seit Jahrzehnten etabliert und auch in Wissenschaftskreisen geläufig sind, wurden in vielen Ländern noch keine Anstrengungen unternommen, diese auch auf den digitalen Bereich zu übertragen (wie z.B. in Österreich).

Dieser Unterschied in der Behandlung analoger und digitaler Ressourcen bringt noch ein weiteres Problem mit sich: Wenige (Geistes-)WissenschafterInnen sind mit den rechtlichen Implikationen ihrer Tätigkeit wirklich vertraut. Viele vertrauen auf langjährige Erfahrungen mit analogen Quellen und übertragen diese auf den Umgang mit digitalen Ressourcen, ohne die aktuelle Gesetzgebung zu rezipieren.

Im Fall von Materialien, die im Eigentum von Universitäten oder Kulturerbeinstitutionen liegen, oder die nach dem Ablauf der Schutzfristen (üblicherweise 70 Jahre nach dem Tod des Urhebers) gemeinfrei sind, bereitet das keine Probleme. Jüngere Quellen – insbesondere im Lichte des wachsenden Interesses an *Big Data* und *Social Media* als Quelle sozial- und kulturwissenschaftlicher Forschung – bringen jedoch gesetzliche Einschränkungen bezüglich ihrer Reproduktion, Verarbeitung und Veröffentlichung mit sich. Ein weiteres Problem zeigt sich anhand der beträchtlichen nationalen Unterschiede in der Handhabung verwahrter Werke.

Häufig nehmen Kulturerbeinstitutionen ein automatisches Recht auf Digitalisate des bei ihnen gelagerten Materials in Anspruch und fordern entsprechende Vergütungen – bisweilen unrechtmäßig, weil z.B. Digitalisierungen nicht durch die Institution selbst durchgeführt wurden oder Rechtsvorschriften jenseits des Urheberrechts (z.B. Landesarchivgesetze) sie zur freien Zurverfügungstellung ihrer Bestände verpflichten.

Zu guter Letzt trägt auch die zunehmende wissenschaftliche Kollaboration über Disziplinen- und Landesgrenzen hinweg zu der ohnehin unklaren Situation bei: Welches Rechtssystem kommt im Fall von digitalen Ressourcen zur Anwendung, die in unterschiedlichen Ländern gehostet werden, welches liegt verteilten elektronischen Publikationen zugrunde?

Dieser Beitrag zeigt zunächst häufige Stolpersteine und juristische Fallen im (geistes-)wissenschaftlichen Forschungsalltag auf und gibt einen Überblick über vorhandene Lösungsansätze und bislang ungelöste Probleme in der Gesetzgebung im deutschsprachigen Raum. Von diesen ausgehend erfolgt ein Vergleich mit internationalen Bestimmungen zu geistigem Eigentum (*Berne Convention for the Protection of Literary and Artistic Works, Trade Related Aspects of Intellectual Property Rights (TRIPS), World Intellectual Property Organization (WIPO) Copyright Treaty*) und den relevanten Richtlinien der EU, mit dem Ziel, im Rahmen bestehender gesetzlicher Bestimmungen und aktueller Entwicklungen (z.B. dem Bekenntnis zur digitalen Lehrmittelfreiheit im deutschen Koalitionsvertrag) realistische *best practice* Beispiele für die freie Nutzung von digitalen Ressourcen in Forschung und Lehre zu definieren.

Literatur

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society.

Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information (amended 2013).

Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights.

Goldstein, Paul et al. (Hrsg.) (2010): International Copyright. Principles, Law, and Practice. Oxford University Press, New York.

Jahnel, Dietmar et al. (Hrsg.) (2012): IT-Recht. Verlag Österreich. Wien.

Kuhlen, Rainer (2008): Erfolgreiches Scheitern – eine Götterdämmerung des Urheberrechts? VWH, Boizenburg.

Schöwerling, Helena (2007): E-Learning und Urheberrecht an Universitäten in Österreich und Deutschland. Verlag Medien und Recht, Wien-München.

Torremans, Paul (Hrsg.) (2007): Copyright Law. A Handbook of Contemporary Research. Edward Elgar Publishing Ltd., Cheltenham.

Poster:

Besonderheiten audiovisueller Forschungsarchive in den „Digital Humanities“

Dr. Jürgen Schöpf, Dipl.-Ing. (FH) Johannes Spitzbart, Dr. Christiane Fennesz-Juhasz
(Phonogrammarchiv der Österreichischen Akademie der Wissenschaften)

Schall- und audiovisuelle Archive bilden ein wesentliches Segment der digitalen Forschungsinfrastruktur. Am Beispiel des Phonogrammarchivs (PhA) der Österreichischen Akademie der Wissenschaften werden die Spezifika von audio-visuellen Forschungsarchiven und ihren Beständen innerhalb sowie ihr Potential für die digitalen Geisteswissenschaften dargestellt.

Das PhA sammelt und bewahrt phonographische und audiovisuelle Original-Quellen, die in Österreich arbeitende Wissenschaftler im Rahmen ihrer Feldforschungen herstellen. Seit seiner Gründung (1899) interdisziplinär definiert wird im PhA bis heute sowohl naturwissenschaftlich (Restauration von Aufnahmen auf analogen Medien) wie geisteswissenschaftlich (ethnologisch, musikologisch, linguistisch) geforscht. Das Phonogrammarchiv arbeitet seit vielen Jahren aktiv an internationalen Standards der Digitalisierung mit (z.B. IASA-TC04). Teile seiner Sammlungen sind im Weltdokumenten-Register „Memory oft the World“ der UNESCO eingetragen. Der Focus auf Ton- und seit 2000 auch Video-Quellen sowie umfangreiche Erfahrungen in deren Digitalisierung und Langzeitarchivierung bringen einige Besonderheiten mit sich, die im Folgenden zusammengefasst und an Beispielen erläutert werden:

- Das audio-visuelle Archiv als digitale Forschungsinfrastruktur umfasst einerseits die Digitalisierung historischer Tonaufnahmen, die Archivierung digital entstandener Forschungs-Rohdaten, die elektronische Verfügbarmachung der audio-visuellen Bestände und deren Langzeitarchivierung; sowie andererseits den Bestandsnachweis mit detaillierten Suchmöglichkeiten in einer in-House Datenbank und einem Online-Katalog. Der Katalog des PhA ist auch über das Dismarc/Europeana-Portal durchsuchbar; mit seinen Beständen, aber auch seiner Fachexpertise bezüglich der Herstellung und Langzeitbewahrung von AV-Quellen wird sich das PhA ab 2014 verstärkt an den Initiativen CLARIN und DARIAH beteiligen.
- Audiovisuelle Forschungsdaten sind Unikate und werden als multidisziplinär auswertbare Primärquellen behandelt. Solche Primärquellen können nicht z.B. durch eine Texttranskription oder eine Schallanalyse ersetzt werden. Sie sollen für zukünftige, heute noch nicht absehbare Fragestellungen zur Verfügung stehen. Das audiovisuelle Dokument ist die Archivalie.
- Eine dazugehörige Quellenkritik muss diese Dokumente kontextualisieren. Dies bedeutet, den technischen, wissenschaftshistorischen wie auch ethnohistorischen Rahmen ihres Zustandekommens zu benennen. Aus diesem Grund veröffentlicht das Phonogrammarchiv eine CD-Reihe als Quellenedition seiner historischen Bestände, in der besonderer Wert auf die Kontextualisierung gelegt wird, um heutigen Forschern diese historischen Kontexte bewusst zu machen. Dies kann als „analoge Quellenkritik“ bezeichnet werden. Für viele historische Bestände gibt es ethische Probleme die bei ihrer Veröffentlichung – sei es als CD-Quellenedition oder im Internet – jeweils berücksichtigt werden müssen.
- Darüber hinaus muss eine „digitale Quellenkritik“ den technischen Bedingungen der Digitalisierung sowie dabei anfallenden Metadaten Rechnung tragen. Die Ent-Medialisierung oder auch Ent-Materialisierung der Inhalte verlangt eine umso akribischere Dokumentation des Digitalisierungsprozesses, insbesondere der technischen und kulturhistorischen Grenzen, die zum Zeitpunkt der Aufnahme Einfluss auf die Inhalte genommen haben (z. B. Kürzung auf Medienlänge, soziale Hierarchien der Beteiligten).

- Langzeitarchivierung erfordert eine regelmäßig wiederkehrende Migration aller Daten auf neue Speichermedien, dies vergrößert in jedem Migrationsschritt die technischen Metadaten. Das Phonogrammarchiv migriert aktuell von der ersten Generation Digital-Video (ab 2000) auf eine zweite Generation. Da bis heute keine Standards für die (Langzeit)Archivierung von Video vorliegen, erzwingt diese Migration auch eine Formatmigration.
- Originalmedien und -Abspielgeräte müssen gepflegt werden. Die kontinuierliche Forschung an analoger Restauration bietet Möglichkeiten, in Zukunft noch mehr Informationen aus historischen analogen Trägern gewinnen zu können (z. B. Auslesen der HF-Bias analoger Tonbänder zur Korrektur von Gleichlaufschwankungen, Rekonditionierung nicht mehr abspielbarer Tonbänder).
- Inhaltsbeschreibende Algorithmen für Audiodaten („MIR“) sind noch in den Anfängen. Ihre Entwicklung erfordert hohen technischen Aufwand, sie sind aber meist nur für sehr spezielle Forschungsfragen nutzbar. (Beispiel 1: MIDI-Enkodierung von Volksliedmelodien innerhalb eines bekannten kulturellen Rahmens. Beispiel 2: Spektrogramm-Editor als langjähriges Desiderat in der systematischen Musikwissenschaft und Musikethnologie. Problem: Forschungsfragen der Geisteswissenschaften lassen sich nicht mit Forschungsfragen der Naturwissenschaften verbinden. Eine Begegnung „auf Augenhöhe“ findet nicht statt).

1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014)

Universität Passau · 25.-28. März 2014 * Abstract für einen Vortrag

Einfluss immersiver Benutzerschnittstellen auf kognitive Prozesse in virtuellen Lernumgebungen

Katharina Schuster, Ursula Bach, Anja Richert, Sabina Jeschke

Einführung

Etwa seit den 1990er Jahren werden Computerspielsimulationen für Lern- und Trainingszwecke entwickelt und eingesetzt, um komplexe technische Systeme oder den Umgang mit ökonomischen (z.B. Unternehmensplanspiel) ökologischen (z.B. Klimaentwicklung) oder sozialen Systemen (z.B. Stadtplanung) zu erlernen (Kerres 2012). Simulationen eignen sich für unterschiedliche Szenarien:

- Visualisierung komplexer theoretische Konstrukte (z.B. der Mathematik), die aus verschiedenen Perspektiven betrachtet und verändert werden können (Jeschke 2004),
- Situationen, in denen Fehlverhalten zu gesundheitlichen Schäden oder schlimmstenfalls dem eigenen Tod bzw. dem Tod anderer Menschen führt, z.B. ein Störfall in einem Atomkraftwerk (Ewert et al. 2012),
- Situationen, die fast nie eintreten und dennoch möglich sind (Jolie et al. 2011), z.B. Evakuierungen bei Massenpaniken,
- Situationen, deren Training mit hohen Kosten verbunden ist, da bspw. teures Equipment zum Einsatz kommt oder hoher Materialverschleiß mit dem Training einhergeht (Ewert et al. 2012),
- Situationen, die in der Vergangenheit oder in der Zukunft liegen, wie z.B. ein Spaziergang durch das alte Pompeji oder auf dem Mars (Schuster et al. 2013),
- Situationen, in denen nicht sichtbare Wirkungszusammenhänge und Zustände sichtbar gemacht werden sollen, z.B. Luftströme in einem Gebäude (Willenbrock 2012).

Für die Anwendung müssen zunächst oben beschriebene Szenarien in virtuellen Umgebungen nachgebaut werden. In Simulationen oder Lernspielen (Serious Games) können die Eigenschaften und Prinzipien mit entsprechenden Inhalten verknüpft und dadurch nachvollziehbar gemacht werden. Ein Nachteil von Simulationen besteht häufig in der Künstlichkeit der Lernerfahrung. Normalerweise interagieren Nutzerinnen und Nutzer mit einer virtuellen Umgebung über einen PC. Das virtuelle Szenario wird auf einem Monitor angezeigt und das Sichtfeld üblicherweise durch eine Tastatur oder eine Maus gesteuert. Interaktionen mit virtuellen Objekten sowie Fortbewegung erfolgen meist über die gleichen Hardware-Schnittstellen. Auf diese Weise überträgt der Nutzer oder die Nutzerin den Kontrollmodus des Computers auf die Aktionen seiner grafischen Repräsentation, dem Avatar.

Natürliche Benutzerschnittstellen für Visualisierung, Navigation und Interaktion können eine authentischere Lernerfahrung als am PC ermöglichen. Bekannte Beispiele sind Flug- und Fahrsimulatoren oder bestimmte Spielkonsolen im Unterhaltungsbereich. Ziel der natürlichen

Benutzerschnittstellen ist es immer, die Mensch-Computer-Interaktion intuitiver zu gestalten, Illusionen zu erzeugen und bestimmte Situationen so realitätsnah wie möglich zu imitieren. Der Nutzer oder der Nutzerin soll den Eindruck haben, regelrecht in die virtuelle Welt eintauchen zu können. In diesem Zusammenhang fällt häufig der Begriff der Immersion. Auch wenn der Begriff höchst heterogen verwendet wird, so kann Immersion in einer ersten Annäherung als „der subjektive Eindruck, dass jemand eine umfassende und realistische Erfahrung macht“ (Dede 2009) definiert werden. Oft ist es das erklärte Ziel von Entwicklern virtueller Welten oder natürlicher Benutzerschnittstellen, den Eindruck des „Eintauchens“ durch bestimmte technische Eigenschaften zu unterstützen. Dementsprechend ist in solchen Fällen von immersiven Lernumgebungen oder immersiven Benutzerschnittstellen die Rede. Aus technischer Sicht benötigt der Nutzer oder die Nutzerin für eine erhöhte Immersion eine nahtlose 3D-Sicht der virtuellen Umwelt. Diese wird häufig durch Head Mounted Displays (HMDs) realisiert. Für eine natürliche Navigation in der virtuellen Umgebung können omnidirektionale Laufbänder verwendet werden, die eine freie und unbegrenzte Bewegung ermöglichen und die nutzende Person doch an einem physisch klar begrenzten Ort lassen. Durch Datenhandschuhe kann der Nutzer bzw. die Nutzerin intuitiv mit der virtuellen Umgebung sowie den Objekten, die sich in ihr befinden, interagieren. Die beschriebenen Komponenten – HMD, omnidirektionales Laufband und der Datenhandschuh – sind im so genannten Virtual Theatre integriert (MSEAB Weibull 2012). Abbildung 1 zeigt die Nutzung des Virtual Theatres sowie ein exemplarisches Anwendungsszenario aus den Ingenieurwissenschaften.

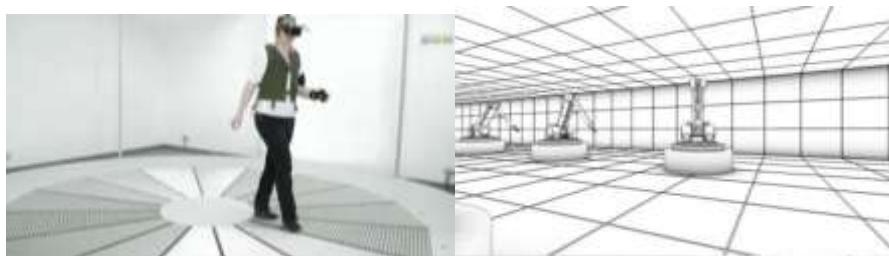


Abbildung 1: Das Virtual Theatre mit Nutzerin und (vereinfachter) virtuellen Umgebung

Bevor immersive Benutzerschnittstellen wie das Virtual Theatre flächendeckend zu Lern- und Trainingszwecken zum Einsatz kommen, sind wissenschaftliche Erkenntnisse über die Wahrnehmung der Rezeptionssituation und über den tatsächlichen Lernerfolg notwendig. Deswegen wird die weitere Entwicklung der Hardware von unterschiedlichen psychologischen Studien begleitet, von denen eine im Folgenden näher beschrieben wird.

Studiendesign

Im Hinblick auf immersive Benutzerschnittstellen ist nach wie vor die Frage offen, unter welchen Bedingungen virtuelle Umgebungen zum besten Lernergebnis führen und noch weiter gefragt, inwiefern Immersion und Lernerfolg miteinander zusammenhängen. Der medienpsychologischen Herangehensweise folgend, ist hierbei die Betrachtung von fünf Faktoren wichtig (Wirth und Hofer 2008):

- Eigenschaften der Mediums
- Eigenschaften der rezipierenden Person
- Die Rezeptionssituation
- Eigenschaften der Technik
- Die Wirkung

Bei der im Vortrag vorgestellten Studie zum Einfluss immersiver Benutzerschnittstellen auf kognitive Prozesse in virtuellen Lernumgebungen wurden die subjektive Erfahrung des Präsenzerlebens (Wirth und Hofer 2008) und des Flows (Rheinberg et al. 2002) als zentrale Kenngrößen für Immersion gemessen. Um weitere Einblicke in die Rezeptionssituation beim Lernen mit immersiven Benutzerschnittstellen zu erhalten, wurden situative Emotionen ebenfalls gemessen. In Anlehnung an Witmer und Singer (1998) folgt die Studie dem Ansatz, dass Faktoren der Erfahrung in einer virtuellen Umgebung in Abhängigkeit individueller Unterschiede, Charakteristika der virtuellen Umgebung wie auch der Hardware des Simulators variieren. Individuelle Unterschiede, Merkmale und Fähigkeiten können in einer bestimmten virtuellen Umgebung z.B. das erlebte Präsenz- und Flow-Empfinden steigern oder mindern. Ebenso können sich verschiedene Charakteristika einer virtuellen Umgebung und der Hardware eines Simulators unterstützend oder beeinträchtigend auf die subjektive Erfahrung in einer Rezeptionssituation auswirken.

Eine der wichtigsten Fragen innerhalb eines Bildungs- und Trainingskontextes ist, ob die Erfahrung in einer virtuellen Umgebung durch immersive Benutzerschnittstellen zu einem besseren Lernergebnis führt als dieselbe Erfahrung über einen PC. Leistungsabfrage ist deshalb eine weitere wichtige Komponente des Studiendesigns. Die erwartete Beziehung zwischen diesen Komponenten sowie den in Abbildung 2 dargestellt.

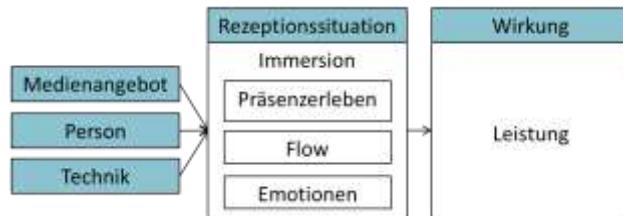


Abbildung 2: Erwarteter Zusammenhang zwischen Medienangebot, Person, Technik, Rezeptionssituation und Wirkung

Für die Studie „Das verrückte Labyrinth“ ($n = 30$) wurden zwei randomisierte Studierendengruppen miteinander verglichen, indem sie in einer Lernsituation unterschiedliche Hardware benutztten. Die Variable der virtuellen Umgebung und die der Aufgabe wurden konstant gehalten: In einem Labyrinth mussten die Versuchspersonen umherlaufen, Objekte finden und sich deren Positionen merken. Dafür hatten sie acht Minuten Zeit. Die Experimentalgruppe absolvierte die Aufgabe im Virtual Theatre, während die Kontrollgruppe dieselbe Aufgabe am Laptop löste. Das Leistungsabfrageszenario wurde wiederum für beide Gruppen konstant gehalten. Hierfür mussten die Versuchspersonen an einem Tablet PC die Objekte per Drag and Drop Steuerung an ihre korrekten Positionen auf einer Karte des leeren Labirynths zuordnen (siehe Abbildung 3).



Abbildung 3: Leistungsabfrage am Tablet PC

Der Tablet PC wurde gewählt, damit sich das Medium der Wissensabfrage sowohl für die Experimental- als auch für die Kontrollgruppe hinreichend vom Medium der Aufgabe unterscheidet. Die Genauigkeit der Position der Objekte und die benötigte Zeit wurden vom Computer automatisch erfasst, die Ergebnisse dienten als numerische Kennwerte für die erbrachte Leistung. Die persönliche Wahrnehmung der Rezeptionssituation wurde mit entsprechenden Skalen per Online-Fragebogen gemessen, der ebenfalls über den Tablet PC ausgefüllt wurde. Die Daten wurden statistisch mit SPSS ausgewertet. Hierfür wurden ANOVAs zwischen den Gruppen gerechnet.

Ergebnisse

Erste Ergebnisse zeigen, dass die Virtual Theatre Bedingung nicht zu mehr Präsenzerleben führt, als die Laptopbedingung. Dies ist wohl durch die Störvariablen zu begründen, wie z.B. lautes Betriebsgeräusch des omnidirektionalen Laufbands, Gewicht des Head Mounted Displays usw. Die Virtual Theatre Bedingung führt jedoch zu mehr negativen Emotionen. Dies könnte an der Kombination des Head Mounted Displays mit realer physischer Fortbewegung liegen. Mit einem HMD kann man als Nutzer oder Nutzerin zwar sehen, wohin man in der virtuellen Umgebung läuft, aber in der physischen Umgebung ist man sozusagen blind. Ergo geht man ein gewisses Risiko ein, da man stolpern oder fallen könnte, insbesondere weil das Laufen auf dem omnidirektionalen Laufband ungewohnt ist. Es ist also ein gewisses Vertrauen in die Technik von Nöten, um dieses Risiko auszublenden. In der aktuellen Stichprobe wurde am Laptop bessere Leistung erbracht als im Virtual Theatre. Dies kann an der höheren kognitiven Beanspruchung liegen, die mit der Nutzung des Virtual Theatres einhergeht. Alle oben erwähnten Prozesse, das Abwägen von Risiko, das Unterscheiden von physischer und virtueller Realität usw. müssen zeitgleich kognitiv verarbeitet werden. Somit bleiben weniger Ressourcen für die Bearbeitung der eigentlichen Aufgabe.

Die Ergebnisse bestätigen die Notwendigkeit, immersive Benutzerschnittstellen hinsichtlich ihres technischen Potenzials immer mit den tatsächlich erlebten Emotionen und Wahrnehmungsparametern der Nutzergruppen abzugleichen. Durch eine genaue Überprüfung des Zusammenhangs zwischen Nutzerin bzw. Nutzer, Hardware, virtueller Umgebung und Lernerfolg ist es möglich, den Anteil der Immersion festzustellen, der durch die Persönlichkeit beeinflusst wird. Dadurch kann wiederum ein Profil erstellt werden, für welche Nutzergruppen immersive Benutzerschnittstellen den größten Mehrwert bieten. Außerdem können so maßgeschneiderte Lernszenarien für unterschiedliche Nutzerprofile entwickelt werden.

Literatur

- Dede, C. (2009): Immersive Interfaces for Engagement and Learning. Science 2 January 2009
- Ewert, D., Schuster, K., Schilberg, D., Jeschke, S. (2012): Intensifying learner's experience by incorporating the virtual theatre into engineering education. In: Proceedings of the IEEE Educon Conference, 13. – 15. März, Berlin
- Jeschke, S.: Mathematik in Virtuellen Wissensräumen – IuK-Strukturen und IT-Technologien in Lehre und Forschung. Dissertation. 2004.
- Jolie, S.; Katzky, U.; Bredl, K.; Kappe, F.; Krause, D. (2011): Simulationen und simulierte Welten. Lernen in immersiven Lernumgebungen. In: Ebner, M.: Schön, S.: Lehrbuch für Lernen und Lernen mit Technologien.
- Kerres, M. (2012): Mediendidaktik. Konzeption und Entwicklung mediengestützter Lernangebote. München: Oldenbourg.
- MSEAB Weibull (2012): The Virtual Theatre. Online im Internet: <http://www.mseab.se/The-Virtual-Theatre.htm> (Zugriff am 15.12.1012)
- Rheinberg, F.; Engeser, S.; Vollmeyer, R. (2002): Measuring components of flow: the Flow-Short-Scale. In: Proceedings of the 1st International Positive Psychology Summit, Washington DC, USA
- Schuster, K., Ewert, D., Johansson, D., Bach, U., Vossen, R., Jeschke, Sabina (2013): Verbesserung der Lernerfahrung durch die Integration des Virtual Theatres in die Ingenieurausbildung In: Tekkaya, A. E.; Jeschke, S.; Petermann, M.; May, D.; Friese, N.; Ernst, C.; Lenz, S.; Müller, K.; Schuster, K.(Hg.): TeachING-LearnING.EU discussions. Innovationen für die Zukunft der Lehre in den Ingenieurwissenschaften.
- Willenbrock, H. (2012): Wie im richtigen Leben. In: Brandeins, 14. Jahrgang, Heft 10.
- Wirth, W.; Hofer, M. (2008): Präsenzerleben. Eine medienpsychologische Modellierung. In: Montage AV. Zeitschrift für Theorie und Geschichte audiovisueller Kommunikation, 17/2/2008, S. 159 – 175.
- Witmer, B.G.; Singer, M.J. (1998): Measuring Presence in Virtual Environments: A Presence Questionnaire, *Presence: Teleoperators and Virtual Environments*, Vol. 7, Nr. 3, S. 225 – 240

Digitale Edition als Methode kunsthistorischer Forschung: Die Werktagebücher von Hartmut Skerbisch

Martina Semlak, Universität Graz

Der Beitrag präsentiert den aktuellen Stand des Dissertationsprojekts der digitalen genetischen Edition der Werktagebücher des österreichischen Konzeptkünstlers Hartmut Skerbisch (1945-2009).

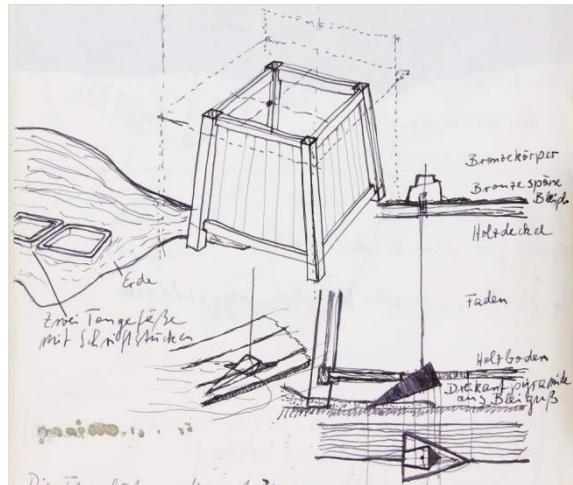
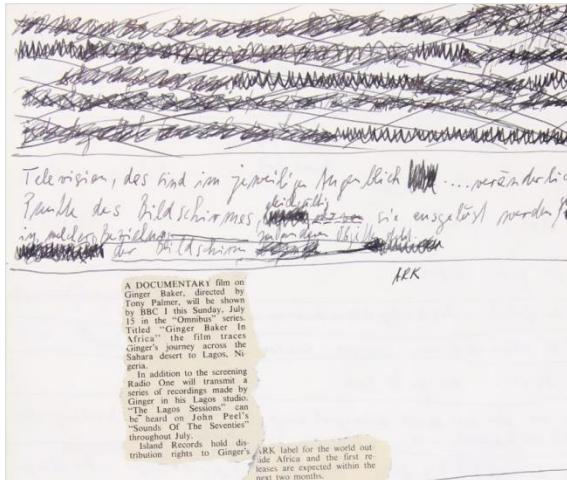
In diesem Projekt wird der Nutzen semantischer Technologien für die Erschließung kunsthistorischen Quellenmaterials untersucht: Die zentrale Frage ist, wie eine digitale, genetische und semantisch angereicherte Edition das Verständnis von Konzepten und Assoziationen des Künstlers im Schaffensprozess und damit letztlich die Rezeption des Kunstwerks unterstützen kann.

Werktagebücher, Skizzenbücher und Notizen von Künstlern geraten immer mehr in den Fokus kunsthistorischer Forschung und nehmen eine wichtige Stellung zur Untersuchung der Werkgenese ein. Diese Quellengattungen gelten als unmittelbare Zeugen des Künstlers und dessen Werkschaffensprozesses, indem sie Einblicke in den Alltag des Künstlers und seine Umgebung gewähren. Die Bedeutung solcher Quellen in der kunsthistorischen Forschung zeigt sich auch am Beispiel der Edition der Skizzenbücher von Max Beckmann und der digitalen Edition der Unterrichtsnotizen zur Form- und Gestaltungslehre von Paul Klee.

Dennoch bleibt die Edition eine in der Kunstgeschichte bisher wenig verbreitete Methode. Eine digitale Edition mit ihren Möglichkeiten zur strikten Trennung von Form und Inhalt und damit Unabhängigkeit von Ausgabeformen, der Verknüpfung mit bereits existierenden Ressourcen und kontrollierten Vokabularen und dem Einsatz von Werkzeugen zur Herstellung von Text-Bild-Beziehungen sowie zur Analyse und Visualisierung des Materials bietet ein breites Spektrum an Möglichkeiten für die kunsthistorische Forschung.

Hartmut Skerbisch war ein österreichischer Konzeptkünstler der sich mit dem Raumbegriff und einer neuen Auffassung des Skulpturbegriffs auseinandersetzte: dabei waren die Beziehung von Objekten zueinander, zum umgebenden Raum und zum Betrachter zentrales Motiv. Texte waren häufig Ausgangspunkt seiner Werke: er zitierte Literaten wie James Joyce, Franz Kafka oder Kathy Acker, befasste sich mit Texten von Musikern wie Lou Reed und reflektierte diese in seinen Werken. Daneben treten Einflüsse aus so weit gestreuten Bereichen wie elektronische Medien, Mathematik, Physik, Philosophie und Anthropologie.

Die 35 Werktagebücher entstanden in einem Zeitraum zwischen 1969 und 2008 und umfassen in etwas 2100 beschriebene Seiten. Es handelt sich dabei um handschriftliche Texte, Kalkulationen und Formeln sowie Skizzen. Die Notizen von Skerbisch folgen keiner linearen Struktur, sondern sind meist fragmentarisch, er verzichtet auf formale Notationen. Passagen wurden übermalt, aus dem Heft herausgerissen, Zeitungsausschnitte eingeklebt, Texte gestrichen oder korrigiert. Diese zunächst zufällig angeordnet erscheinenden Elemente wirken jedoch bei näherer Betrachtung bewusst arrangiert: ihre Anordnung verleiht dem Text eine zusätzliche Bedeutung, sie werden selbst zur künstlerischen Komposition. Dieses Merkmal verlangt zusätzlich zur werkorientierten nach einer dokumentorientierten Betrachtung. Zur digitalen Kodierung von Texten hat sich die TEI zu einem de-facto Standard entwickelt. Im vorzustellenden Projekt werden die aus einer stark textorientierten Tradition stammenden Richtlinien auf ihre Anwendbarkeit im kunsthistorischen Kontext geprüft.



Um sich den Assoziationsprozessen des Künstlers anzunähern, werden unterschiedliche Herangehensweisen eingesetzt:

1. Mit der Arbeit soll untersucht werden, inwieweit die traditionelle philologische Edition für kunsthistorische Fragestellungen eingesetzt werden kann und wo hier die Grenzen liegen. Dies betrifft vor allem den Umgang mit grafischen Elementen, die hier eine besondere Stellung einnehmen.
2. Der Prozess des Schreibens vollzieht sich in Raum und Zeit: Gedanken treten hervor, sie verändern sich über die Jahre, manifestieren sich in einem Werk oder geraten wieder aus dem Blickfeld. Die Entwicklung des Textes über die Zeit ist daher ein relevanter Aspekt um die Werkgenese nachzuzeichnen. Besonders moderne Manuskripte, mit fragmentarischen und flüchtigen Notizen, liegen nicht als abgeschlossene Dokumente, sondern vielmehr als „unfertige“ Entwürfe vor. Zur Rekonstruktion des Schreibprozesses bietet sich die Methode der genetischen Edition an: diese fokussiert einen dokumentzentrierten Ansatz und berücksichtigt Veränderungen des Textes wie Korrekturen, Streichungen und Hinzufügungen durch den Autor. Die Kodierung solcher Phänomene stellt eine editorische Herausforderung dar, für die die TEI ein Instrumentarium bietet. Die von der Arbeitsgruppe „Genetic Edition“ vorgeschlagenen Elemente und Attribute werden in der Praxis erprobt.
3. Ein weiterer Schwerpunkt liegt auf der Untersuchung der Anwendbarkeit semantischer Technologien um die kulturellen und intellektuellen Grundlagen des Kreativprozesses zu erforschen. Dazu werden Referenzen im Text mit Konzepten wie Personen, literarische Werke, Musik oder Orte verknüpft. Unter Nutzung kontrollierter Vokabulare (z.B. GND, VIAF, GeoNames) wird eine projektinterne erweiterbare Ontologie erstellt. Die semantischen Beziehungen zwischen den so annotierten Konzepten können abgefragt und visualisiert werden und so Zusammenhänge sichtbar machen, die sich dem Leser üblicherweise nicht auf den ersten Blick erschließen.

Durch die Aggregation der Methoden sollen

- a) die fragmentarischen und thematisch weit reichenden Einträge der Notizbücher in eine inhaltliche Reihenfolge gebracht werden,
- b) Relationen von Konzepten visualisiert werden, um die facettenreichen und rhizomartigen Einflüsse auf Skerbischs künstlerisches Werk zu dokumentieren,

- c) eine Verbindung zwischen den Tagebucheintragungen und den von Skerbisch realisierten Werken hergestellt werden und schließlich
- d) der kreative Prozess des Künstlers nachgezeichnet werden, um die Rezeption seines Werkes zu unterstützen.

Literatur

- Allemang, D. und Hendler, J. (2011). Semantic Web for the Working Ontologist. Effective Modeling in RDFS and OWL.
- Brüning, G., Henzel, K. und Pravida, D. (2013). Multiple Encoding in Genetic Editions: The Case of 'Faust'. In: Journal of the Text Encoding Initiative. <http://jtei.revues.org/697>, 16. Dezember 2013.
- Burnard, L., Jannidis, F., Pierazzo, E. und Rehbein, M. (2008-2013). An Encoding Model for Genetic Editions. <http://www.tei-c.org/Activities/Council/Working/tcw19.html>, 16. Dezember 2013.
- Fenz, W. (1994). Hartmut Skerbisch. Werkauswahl 1969-1994.
- Glasmeier, M. (1994). Die Bücher der Künstler. Publikationen und Editionen seit den sechziger Jahren in Deutschland.
- Pierazzo, E. (2009). Digital Genetic Editions. The Encoding of Time in Manuscript Transcription. In: Deegan, M. und Sutherland, K. (Hrsg.), Text Editing, Print and the Digital World, 169-186.
- Robinson, P. (2013). Towards a Theory of Digital Editions. In: Variants – Journal of the European Society for Textual Scholarship, 10. 105-131.
- Sahle, P. (2013): Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik. In: Schriften des Instituts für Dokumentologie und Editorik, 8.
- Text Encoding Initiative, TEI P5 Guidelines, <http://www.tei-c.org/Guidelines/P5/>, 16. Dezember 2013.
- Zeiller, C. (2010). Max Beckmann. Die Skizzenbücher. Ein kritischer Katalog. 2 Bände.
- Zentrum Paul Klee (2011), Paul Klee – Bildnerische Form- und Gestaltungslehre. <http://www.kleegestaltungslehre.zpk.org>, 16. Dezember 2013.



DHd 2014 in Passau
Digital Humanities – methodischer Brückenschlag
oder "feindliche Übernahme"?

Posterpräsentation

Eine digitale Ausgabe des „Welschen Gastes“ als Chance für neue Analyse- und Visualisierungsmethoden

Die seit 2011 im Rahmen des Heidelberger Sonderforschungsbereichs 933 „Materiale Textkulturen“ entstehende digitale Neuausgabe des mittelhochdeutschen Text-Bild-Gedichts „Welscher Gast“ Thomasins von Zerklaere kodiert Volltranskriptionen der Handschriften und editorisch hergestellte Texte im XML/TEI-Format und legt damit verknüpfte Bildannotationen in einer relationalen Datenbank ab. Das Ziel des Projekts sind eine Online-Edition des gesamten textuellen und bildlichen Materials, welche die komplexen Text-Bild-Beziehungen im Werk dokumentiert und mit mächtigen Visualisierungsmechanismen anschaulich präsentiert, sowie mehrere Buchausgaben mit jeweils abgestufter Komplexität und unterschiedlichem Zielpublikum. Die Text- und Bilddaten sollen leicht vergleichbar und durchsuchbar gemacht werden.

Für die stemmatologische Analyse der transkribierten, lemmatisierten und alinierten Textdaten machen wir uns phylogenetische Software aus der Bioinformatik zunutze, die es erlaubt, mutmaßliche Verwandschaftsbeziehungen zwischen den überlieferten Handschriften in gewurzelten Baumgraphen und ungewurzelten Netzwerken darzustellen. Die so gewonnenen Erkenntnisse setzen wir in der digitalen Textausgabe um, um nicht nur herkömmliche Synopsen zu generieren, sondern auch neue Formen des kritischen Apparats zu erproben.

Eine Funktion, die wir als „Baumapparat“ bezeichnen, wird dem Benutzer, der entweder den kritisch hergestellten Text oder eine Handschriftentranskription liest, beim Überfahren eines Wortes mit der Maus ein dynamisch generiertes Baumdiagramm zeigen, das dem mutmaßlichen Handschriftenstemma für die jeweilige Textstelle entspricht und an dessen „Ästen“ die Lesarten fürs jeweils ausgewählte Wort sichtbar gemacht werden. Durch farbige Unterlegung wird den angezeigten Lesarten zudem deren semantische Relevanz zugewiesen. Auf diese Weise wird der Benutzer die angezeigten Lesarten unmittelbar und intuitiv ins Stemma einordnen können.

Einen anderen Visualisierungsansatz erproben wir bei der stilometrischen Wortschatz- und Reimanalyse. Die im Text am häufigsten vorkommenden Wort- und Reimkombinationen stellen wir in einem Netzwerk dar, das einen schnellen Eindruck über die sprachlichen Eigenschaften des Werkes vermittelt.

Die in den Handschriften enthaltenen Miniaturzeichnungen und -malereien bilden einen festen Bildzyklus mit einem Bestand von etwa 120 Motiven. Diese Illustrationen enthalten in der Regel allegorische Figuren, die mit Beischriften und Spruchbändern versehen sind. Die Auszeichnung der Bildzonen mit Figuren, Beischriften und Spruchbändern erfolgt in einem browserbasierten graphischen Bildeditor, der die gewonnenen Koordinaten in einer relationa-

len Datenbank speichert. Die TEI-konforme Transkription der Texte in den Bildern wird ebenfalls in dieser Datenbank abgelegt, soll aber später zu Archivierungszwecken als TEI-Dokument exportiert werden. Die verschiedenen Realisierungen einzelner Motive werden in der Datenbank diesen abstrakten Motiven zugeordnet. Ein entsprechendes Alignement erfolgt ebenfalls auf der Ebene der Bild- bzw. Motivkomponenten. Somit werden Visualisierungen möglich, die verschiedene Varianten desselben Motivs nebeneinander stellen und etwa beim Überfahren einer Bildkomponente (z.B. eines Spruchbands) mit der Maus deren jeweilige Pendants in anderen Handschriften graphisch hervorheben und die darin enthaltenen Texte anzeigen.

Die geplante Text-Bild-Ausgabe soll neben verschiedenen Möglichkeiten der Textdarstellung (Synopsen, dynamische Apparate, flexibler Grad der sprachlichen Normalisierung, Verlinkung mit digitalen Wörterbüchern) und der skizzierten Präsentation von Illustrationen insbesondere das Zusammenspiel von Text und Bild in den Handschriften des „Welschen Gastes“ analysieren, dokumentieren und visuell erlebbar machen. Zu diesem Zweck werden inhaltliche und physische Bezüge zwischen einzelnen Textpassagen und den dazugehörigen Illustrationen festgehalten. Dadurch soll es möglich sein, von der Textedition schnell zum entsprechenden Bildmaterial zu gelangen und umgekehrt. Selbstverständlich wird es auch möglich sein, Texte und Bilder im Hinblick auf ihre Platzierung auf einer Handschriftenseite zu betrachten.

Ein für die zweite Projektphase (ab 2015) geplanter Text- und Bildkommentar soll mit der Online-Ausgabe dynamisch verknüpft werden.

Es ist unser erklärtes Ziel, die digitale Text-Bild-Ausgabe der Öffentlichkeit im Open Access verfügbar zu machen und in Zusammenarbeit mit langfristig finanzierten öffentlichen Institutionen (z.B. UB Heidelberg) für deren dauerhafte Zugänglichkeit und Archivierung zu sorgen.

Martin Stark, Universität Hamburg

Michael Kronenwett, Universität Trier

Digital Humanities und Digitale Netzwerkkarten: VennMaker als Softwarewerkzeug für die Geisteswissenschaften

Dieser Vortrag stellt das Softwarewerkzeug VennMaker vor und diskutiert dessen Einsatzmöglichkeiten zur Unterstützung und Bereicherung geisteswissenschaftlicher Forschungen anhand von ausgesuchten Fallbeispielen. Dabei argumentiert der Vortrag in Richtung einer Sichtweise der Digital Humanities als eines „methodischen Brückenschlags“ zwischen den Geisteswissenschaften und der Informatik. Zusätzlich wird der analytische Mehrwert digitaler Werkzeuge für die Geisteswissenschaften diskutiert.

Die Software VennMaker ist ein Werkzeug zur interaktiven und grafischen Erhebung und visuellen Kommunikation und Validierung von sozialen Netzwerken. Seit dem ersten Release der Software im Jahre 2010 wurde das Tool durch ein transdisziplinäres Entwicklerteam, bestehend unter anderem aus Informatikern und Geisteswissenschaftlern, im Rahmen des Forschungsclusters „Gesellschaftliche Abhängigkeiten und Soziale Netzwerke“ der Universitäten Trier und Mainz kontinuierlich mit Hilfe des Feedbacks der internationalen Anwendercommunity weiterentwickelt und gerade auch den geisteswissenschaftlichen Bedürfnissen weiter angepasst. Forschende visualisieren das zu untersuchende Netzwerk schon während des Erstellungsprozesses, indem sie repräsentierende Symbole und Linien in die sogenannten digitalen Netzwerkkarten eintragen. Ein Vorteil dieser Art von Netzwerkkarten sind die vielfältigen Möglichkeiten der Repräsentation und Speicherung von Netzwerkinformationen. Im Vergleich zu eher paper-and-pencil orientierten Tools der qualitativen oder partizipativen Netzwerkforschung können die Größe, Farben und Formen der Netzwerkrepräsentationen leichter modifiziert werden. Zusätzliche grafische Elemente, wie konzentrische Kreise, Sektoren und Tortendiagramme erlauben die Aufnahme und Visualisierung weiterer Daten. Diese visuellen Elemente helfen zudem die Netzwerkdarstellung zu strukturieren und zu standardisieren. Des Weiteren können geografische Karten den Netzwerkdarstellungen hinterlegt werden, dies erlaubt die weitere Akzentuierung der räumlichen Aspekte von Netzwerken. Durch diese visuelle Strukturierung und Standardisierung können unterschiedliche qualitativ erhobene Netzwerke softwaregestützt verglichen und in Interviews oder mit anderen Quellen validiert werden. Zusätzlich unterstützt diese digital abgesicherte Vorgehensweise auch die Untersuchung der zeitlichen Veränderungen von Netzwerkstrukturen.

Ein weiteres Ziel der Software ist es, den Prozess der Codierung, Visualisierung und Analyse sozialer Netzwerke schneller und einfacher zu machen. Traditionelle mehr sozialwissenschaftlich und informatisch orientierte Softwarelösungen im Bereich der quantitativen Netzwerkforschung erfordern das mehr oder minder umständliche Eingeben der relationalen Daten in standardisierter Form in Datenbanken und das Erstellen von Datenmatrizen bevor sie in der Lage sind Netzwerkvisualisierungen zu erstellen. VennMaker

dreht diesen Prozess der Datenerhebung gewissermaßen um. Während die Forschenden ihre Netzwerke mit Hilfe des Tools visualisieren, werden die relationalen Daten im Hintergrund generiert, vorgefertigte Datenmatrizen werden somit nicht länger benötigt und der Forschungsprozess beschleunigt. Dadurch eignet sich das Programm gerade für Forscher aus dem Bereich der Geisteswissenschaften, denen oftmals ein entsprechendes umfassendes Training in den formalen Methoden der Informatik und den Sozialwissenschaften fehlt und für die die traditionellen quantitativen Herangehensweisen an die Thematik der Sozialen Netzwerke eher gewöhnungsbedürftig, wenn nicht gar abschreckend sind. Der VennMaker unterstützt somit die Triangulation von qualitativen und quantitativen Methoden im Rahmen der Netzwerkforschung und kann zudem innerhalb des Forschungsprozesses als heuristisches Tool dienen. Dieses stellt unseres Wissens nach ein Alleinstellungsmerkmal der Software dar.

Während die Forschenden ihre Netzwerke erstellen, kalkuliert VennMaker einige basale Netzwerkmaßzahlen im Hintergrund und stellt diese zur Unterstützung des interaktiven Erstellungsprozesses zur Verfügung. Zusätzlich besteht die Möglichkeit, die visuell generierten Daten in zusätzlichen Arbeitsschritten weiter zu standardisieren oder über zur Verfügung gestellte Schnittstellen als Bilddateien, Tabellen oder Matrizen in andere Programme, welche weitergehende Analysen ermöglichen, zu exportieren. Des Weiteren können zusätzliche externe Informationen zu den Akteuren und deren Beziehungen über das Einbinden von URLs verlinkt werden. Den Abschluss des Vortrages bildet ein Überblick über die aktuellen und geplanten Entwicklungsschritte der Software. Zum einen werden die Optionen zum „freien“ Zeichen von Netzwerken erweitert. Zum anderen werden die Möglichkeiten der visuellen Analyse der temporalen und räumlichen Aspekte von Netzwerkstrukturen massiv ausgebaut. Auch diese Weiterentwicklungen sollen zur Diskussion gestellt werden.

GAMS: Geisteswissenschaftliches Asset Management System

Funktionalitäten und Prozesse eines FEDORA-basierten Digitalen Archivs

Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities (ZIM-ACDH)

Einleitung

Moderne Infrastrukturen für die Verwaltung und Bereitstellung von geisteswissenschaftlichen Daten stehen in einem Spannungsfeld: Einerseits soll die Nachhaltigkeit und Verfügbarkeit im Sinne der Langzeitarchivierung gewährleistet werden, andererseits stellt sich die Erfordernis einer flexiblen und individualisierbaren Nutzung vorgehaltener Inhalte. Mit dem auf FEDORA basierenden Geisteswissenschaftlichen Asset Management System (GAMS) und dem am ZIM-ACDH dafür entwickelten Client *Cirilo* wird der Archivierungsaspekt mit einer Präsentations- und Managementschicht verbunden, sodass idealerweise alle genannten Aufgaben von einer gemeinsamen Infrastruktur erfüllt werden können. Hier handelt es sich um ein institutionelles Repository, das eine Vielzahl von Digitalen Editionen, Bild- und Quellsammlungen enthält.

Die Trennung von Inhalt und Präsentation als grundlegendes Merkmal XML-basierter Textformate impliziert einerseits einen hohen Grad an Flexibilität bei der Analyse und der Transformation so vorliegender (Text-)Daten in unterschiedliche Darstellungs- und Präsentationsformen, erfordert gleichzeitig aber auch darauf abgestimmte Verarbeitungsworkflows im täglichen Umgang mit diesen Daten.

In der Definition des Open Archival Information System Reference Model (OAIS-Referenzmodell) bemüht man sich um eine Standardisierung von Workflows in Digitalen Archiven. Es wurde 2002 als ISO-Norm akzeptiert und hat sich seither breit durchgesetzt. Dabei gibt das Referenzmodell lediglich einen formal-strukturellen Rahmen für das Design von Digitalen Archiven vor und definiert einen offenen Standard für ein dynamisches, erweiterbares Archivinformationssystem mit dem Anspruch auf Allgemeingültigkeit. Es fokussiert auf die nachhaltige Nutzbarmachung digital vorliegender Bestände und dies unter der Berücksichtigung sich ständig verändernder Technologien. In GAMS wurde versucht, zentrale Aspekte dieses Modells zu realisieren.

GAMS

GAMS ist ein Asset Management System zur Verwaltung, Publikation und Langzeitarchivierung digitaler Ressourcen. Mit seinem objektorientierten Ansatz bietet das Open Source Projekt FEDORA einen geeigneten Rahmen für die Implementierung des OAIS-Referenzmodells im Rahmen von GAMS: Ein flexibles, erweiterbares Framework zur

Speicherung, zum Management und zur Dissemination von komplexen digitalen Objekten, das die Versionierung der Inhalte und eine METS-basierte Serialisierung der Objekte für die Langzeitarchivierung ermöglicht. Das der FEDORA-Architektur zugrundeliegende Datenmodell unterstützt die enge Verknüpfung von Inhalt und Metadaten und ermöglicht es so, zusammengehörige Informationen auf einfache und nachvollziehbare Weise gemeinsam zu verwalten. FEDORA bringt dabei eine Vielzahl von nützlichen Merkmalen mit, als Beispiel ist hierbei die standardisierte OAI-PMH-Schnittstelle zu nennen.

In der einfachsten Form stellt eine FEDORA-Objektinstanz lediglich die Aggregation einzelner in einem Klassenmodell beschriebener Inhalte dar und bietet somit eine Speicherinfrastruktur in der zusammengehörige Daten gemeinsam verwaltet werden können: Ein TEI-Dokument und XSL-Stylesheets zur Erzeugung von Präsentationsformen (Analyseansichten, Fassungssynopsen u.Ä.) des Objektinhaltes, ein RDF-Metadatensatz zur Beschreibung semantischer Relationen des Objektes, Bild- und Bildmetadaten etwa der Faksimiles einer im TEI-Dokument transkribierten Handschrift sowie PREMIS-basierte Langzeitarchivierungsmetadaten, die Eigenschaften des Lebenszyklus eines Objektes beschreiben u.v.m. In einer funktionalen Sichtweise beschreibt das FEDORA-Objektmodell aber auch (webservicebasierte) programmgesteuerte Abläufe, die auf den Daten der Objekte operieren: Etwa eine XSLT-Transformation zur dynamischen Erzeugung einer bestimmten Objektansicht beim Zugriff auf das Objekt oder die automatisierte Extraktion semantischer Relationen aus einem TEI-Dokument beim Upload einer Datei.

Cirilo Client

Cirilo ist eine Java-Applikation, die über das Management API (API-M) auf das FEDORA-Repository zugreift und die vor allem für Massenoperationen wie Ingestprozesse oder Ersetzungsvorgänge entsprechende Funktionalitäten mitbringt. Der Client unterstützt den Ingest sowohl aus dem Dateisystem, einer eXist Datenbank oder einer Excel-Tabelle. Während des Ingestprozesses werden Metadaten aber auch semantische Informationen regelbasiert und automatisch aus dem Inhalt extrahiert und in die neu erzeugten Objekte übernommen (zum Beispiel im Dublin Core Format).

Über das Design der oben vorgestellten Inhaltsmodelle (Objektklassen) können in einem FEDORA-Repository komplexe (Objekt-)Klassenhierarchien konstruiert werden. Inhaltsmodelle beschreiben dabei nicht nur die inhaltliche Struktur von digitalen Objekten, sondern können über WSDL (Web Service Description Language) auch so genannte Disseminatoren an die Daten eines Archivobjektes binden. Damit ist es möglich Services mit Datenströmen im Objekt zu verbinden und dynamische Abläufe (z.B. Migrationsvorgänge, XSLT-Transformationen) sowie eventgesteuerte Workflows (z.B. eine Textnormalisierung im Uploadvorgang eines TEI-Dokumentes) zu beschreiben und in die Objektlogik zu verpacken. Der Vorteil dieses Konzepts liegt darin, dass komplexe Datenquellen und Workflows einfach abgewickelt werden können.

Derzeit bietet der Client mehrere Inhaltsmodelle für spezifische Zwecke an, besonders umfangreich ist das TEI-Modell. Der Ingest von TEI-Objekten kann flexibel konfiguriert werden: semantische Informationen können extrahiert werden, referenzierte Bilder werden zum Objekt hinzugefügt oder mit den Textdaten in Verbindung stehende Ontologiekonzepte werden aufgelöst. Ein Abfrageobjekt ermöglicht Suchen mit

spezifischen Parametern in dem in FEDORA integrierten Mulgara Triplestore. Mit Hilfe von Ontologie- und Abfrageobjekten können dynamische Register erstellt werden. Zur Strukturierung des Bestandes können Kollektionen und hierarchische Sammlungen angelegt werden. Einige Inhaltsmodelle sind für spezifische Primärquellen optimiert, beispielsweise METS/MODS, HTML, PDF, BibTeX oder externe Ressourcen, die über URL erreichbar sind. Ein Modell für linguistische Primärquellen wird derzeit in Kooperation mit dem ICLTT der ÖAW entwickelt. Zusätzlich wird getestet, wie kontrollierte Vokabulareien und Thesauri wie geonames.org sinnvoll in die Infrastruktur integriert werden können.

Den Einsatzmöglichkeiten von Disseminatoren im Design von Objektklassen sind eigentlich keine Grenzen gesetzt, somit können über Webservices einfach vorhandene Werkzeuge an Objektinhalte gebunden werden. Systemadministratoren erhalten damit ein höchst flexibles Werkzeug zum Design und zur Parametrierung eines Digitalen Archives. Aus einer NutzerInnenperspektive ist lediglich die Instanziierung einer Objektklasse im Repository nötig, um die in einer Objektklasse gekapselte Funktionalität zur Verfügung zu haben. Das METS/MODS-Objekt ist für die Anzeige im DFG-Viewer optimiert, TEI-Objekte können als Grundlage für die Anwendung der Voyant Tools oder der Versioning Machine dienen; Inhalte, die placeName-Elemente enthalten, können auf eine Google Map projiziert oder im Geo-Browser dargestellt werden. Häufig kommen projektspezifische XSL-Stylesheets zum Einsatz, die die Objekte in eine gewünschte Anzeigeform bringen.

Der Client Cirilo wird noch 2014 als österreichischer Beitrag zu DARIAH-EU unter einer Open-Source Lizenz zusammen mit einer umfangreichen Dokumentation zur Verfügung stehen.

Ablauf

Der Workshop ermöglicht es den TeilnehmerInnen, den Client Cirilo auf einer Instanz des GAMS Repositoriums zu testen. Dabei wird zwischen vortragendenzentrierten Impulseinheiten und praktischen Übungen der TeilnehmerInnen abgewechselt: Zunächst werden bestimmte Arbeitsabläufe und Funktionalitäten anhand von konkreten, in Projekten mit unterschiedlichen Quellenmaterialien und Forschungsinteressen bereits umgesetzten Lösungen demonstriert und erläutert. Danach können die TeilnehmerInnen, entweder anhand eigener Projektdaten oder mittels vorbereiteter Beispieldateien diese in der Testumgebung erproben.

Die Vortragenden stehen während des gesamten Workshops für Fragen zur Verfügung und unterstützen die TeilnehmerInnen während der praktischen Übungen.

Literatur

DARIAH-EU: www.dariah.eu [2013-10-28]

DFG-Viewer: <http://dfg-viewer.de/ueber-das-projekt/> [2013-10-28]

FEDORA Commons: <http://www.fedora-commons.org/> [2013-10-28]

Geisteswissenschaftliches Asset Management System: <http://gams.uni-graz.at/> [2013-10-28]

Geo-Browser: <http://geobrowser.de.dariah.eu/> [2013-10-28]

Google Maps: <https://maps.google.at/> [2013-10-28]

Carl Lagoze, Sandy Payette, Edwin Shin, Chris Wilper, Fedora. An Architecture for Complex Objects and their Relationships. 2005. <http://arxiv.org/ftp/cs/papers/0501/0501012.pdf> [2013-10-28]

Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012 <http://public.ccsds.org/publications/archive/650x0m2.pdf> [2013-12-02].

Johannes Stigler. Think global, act local. Reale Probleme und virtuelle Lösungen. Eine Bestandsaufnahme anlässlich 50 Jahre Österreichische Mediathek und des UNESCO-World-Day for Audiovisual Heritage 2010. Hg. von G. Fröschl & R. Hubert & E. Murlasits & S. Steinlechner, LIT Verlag Wien 2012, S. 113-126.

Johannes Stigler (gemeinsam mit W. Hofmeister). Edition als Interface. Möglichkeiten der Semantisierung und Kontextualisierung von domänen spezifischen Fachwissen in einem Digitalen Archiv am Beispiel der XML-basierten Augenfassung zur Hugo von Montfort-Edition, in Editio 24/2010, S. 39-56.

Versioning Machine: <http://v-machine.org/> [2013-10-28]

Voyant Tools: <http://voyant-tools.org/> [2013-10-28]

Prof. Dr. Uta Störl, Hochschule Darmstadt, Fachbereich Informatik, uta.stoerl@h-da.de
Prof. Dr. Hartmut Vinçon, Hochschule Darmstadt, Editions- und Forschungsstelle Frank Wedekind, hartmut.vincon@h-da.de

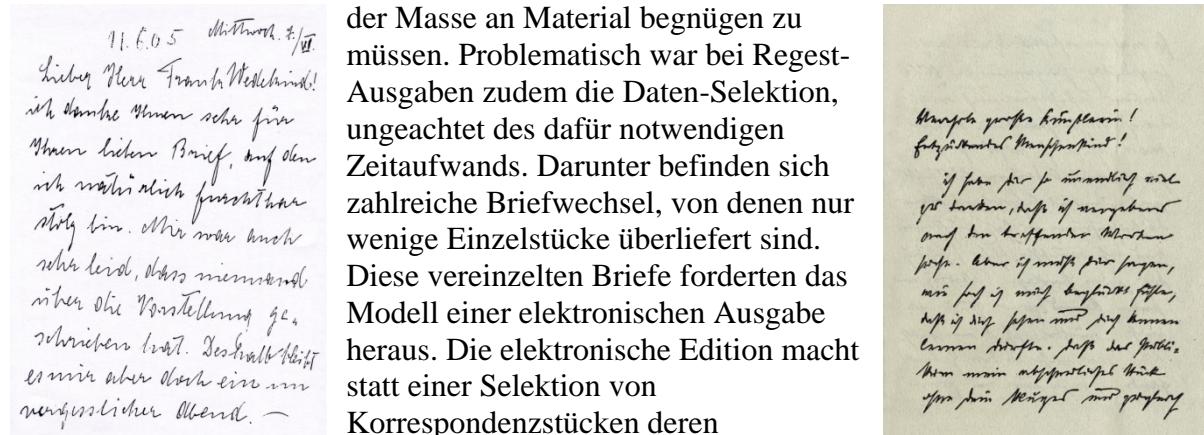
Kooperatives Forschungsprojekt

Online-Brief-Datenbank. Ein Beispiel für disziplinspezifische Anwendungen

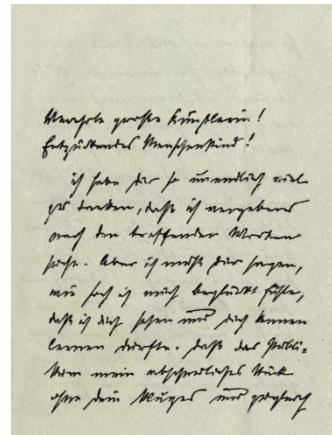
Als kooperatives Forschungsprojekt zwischen dem Fachbereich Informatik und der Editions- und Forschungsstelle Frank Wedekind (EFW) wird an der Hochschule Darmstadt eine „Online-Volltext-Datenbank“ zur Archivierung und Kommentierung von Brief-Korpora entwickelt. Zur Erprobung stehen dafür ca. 3.300 Korrespondenzstücke (Briefe, Postkarten, Telegramme etc.) von und an Frank Wedekind zur Verfügung.

Editionswissenschaftliche Aspekte

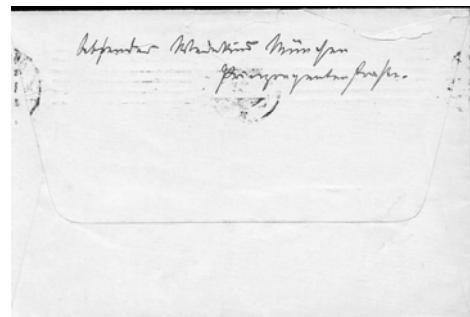
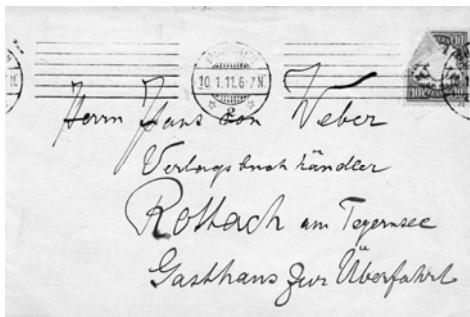
Für die Edition großer Brief-Korpora – wie in diesem Fall – eignet sich die elektronische Edition hervorragend, statt sich – wie früher – mit Brief-Regesten angesichts



der Masse an Material begnügen zu müssen. Problematisch war bei Regest-Ausgaben zudem die Daten-Selektion, ungeachtet des dafür notwendigen Zeitaufwands. Darunter befinden sich zahlreiche Briefwechsel, von denen nur wenige Einzelstücke überliefert sind. Diese vereinzelten Briefe forderten das Modell einer elektronischen Ausgabe heraus. Die elektronische Edition macht statt einer Selektion von Korrespondenzstücken deren



vollständige Publikation möglich, insbesondere auch die weniger vorhandener Einzelstücke bei zahlreichen Briefwechseln, die bei Printeditionen oft unberücksichtigt blieben. Zudem können im Prinzip jederzeit Updates erstellt werden. Ebenso sprechen für die elektronische Edition ihre vielfältigen Darstellungs- und Recherchemöglichkeiten. Schließlich besteht die



Möglichkeit, die auf einer Datenbank gespeicherten Dokumente in verschiedenen Medien zu publizieren, nicht zuletzt auch in einer Print-Edition, was z.B. für einzelne Briefwechsel wünschenswert sein kann. Die Frage der Nachhaltigkeit der elektronischen Editionen heute bleibt freilich bestehen. Wir können nicht prognostizieren, wie z.B. in fünfzig Jahren sich die elektronische Datenverarbeitung weiter entwickelt hat, auch wenn wir heute von XML als langlebigem und plattformunabhängigem Speicherformat ausgehen. Der durch die Informatisierung ausgelöste Medienwandel macht auf den prozessualen Charakter, der Medien per se eignet, deutlicher aufmerksam, als es je durch den Buchdruck der Fall war.

Im Einzelnen sind für die „Eingabe“ und die „Ausgabe“ einer Online-Brief-Datenbank folgende Hauptkategorien, denen zahlreiche weitere untergeordnet sind, berücksichtigt: Eigenschaften des Korrespondenzstücks (Sender/Empfänger/Textsorte des Korrespondenzstücks/Blatt- und Seitenzahl/Materialität des Dokuments); Datum/Ort; Zustellweg; Inhalt (Dokumenttext); Kuvert; Beilagen; Fassungen; Faksimiles; Erstdruck; Standort. Es geht jedoch nicht nur um die Erfassung von Briefmaterialien, vielmehr wird auch die briefspezifische Relation von Textinhalt und Textform editorisch repräsentiert und kommentiert.

Semantische Aspekte

Semantischer Dreh- und Angelpunkt sind selbstverständlich die Kommentare zu den Brieftexten, die durch Einzelstellenkommentare und Biographien – jeweils nach Person, Ort, Örtlichkeit, Werk und Ereignis – ergänzt und miteinander vernetzt sind. Erst eine werkgeschichtliche und biographische Beziehungen transzendierende Betrachtungsweise, welche jene im Kontext sozialer und kultureller Bewegungen reflektiert, hilft, übergreifende Zusammenhänge zu erschließen, und lässt Korrespondenz als Dokumente einer Kulturgeschichte begreifen. Dank der jetzt vollständig vorliegenden historisch-kritischen Print-Edition, welche die Werke Wedekinds im kulturgeschichtlichen Kontext der Zeit verankert, lassen sich parallel dazu in der Online-Briefausgabe vielfältige semantische Beziehungen zwischen Werk- und Briefcorpus herstellen. Diese Kontextualisierung entspricht hervorragend dem Phänomen „Brief“.

Neben der Editionsgeschichte der Werke und Briefe Frank Wedekinds ist das mit ihr eng verknüpfte Desiderat einer kulturwissenschaftlichen Analyse seiner sehr vielfältigen Korrespondenz eine weitere spezifische Voraussetzung für die Edition. Erst eine werkgeschichtliche und biographische Beziehungen transzendierende Betrachtungsweise, welche jene im Kontext sozialer und kultureller Bewegungen reflektiert, hilft, übergreifende Zusammenhänge zu erschließen, und lässt Korrespondenz als Dokumente einer Kulturgeschichte begreifen. Zwar sind in den letzten Jahren erhellende Studien über einzelne Briefwechsel entstanden, und nicht zuletzt muss die in den 20er Jahren von Fritz Strich herausgegebene erste Briefsammlung, die ausschließlich Briefe von Wedekind enthält, erwähnt werden. Über das Briefwerk in Gänze gibt es aber bislang keine wissenschaftlichen Veröffentlichungen, obwohl sich gerade im Kaleidoskop der zahlreichen Korrespondenzpartner/innen die Dynamik der kulturellen Beziehungen jener Zeit zwischen Politik, Justiz und Kunst wahrnehmen lässt.

Das Faszinierende an der Textsorte Brief ist ihr permanentes Oszillieren zwischen privaten und öffentlichen Diskursen, die sich in den jeweiligen Epochen natürlich unterschiedlich formieren. Zwar ist eine Trennung von persönlich-biographischen und literarischen Textgattungen aus heutiger kulturwissenschaftlicher Perspektive ohnehin nicht mehr aktuell, aber gerade in der zurückliegenden germanistischen Einschätzung der Korrespondenz Wedekinds lässt sich die Schwierigkeit erkennen, mit dieser Schnittstelle persönlicher und gesellschaftlicher Kommunikationsformen souverän umzugehen. Die an ihn gerichteten Briefe hat er selbst sorgfältig gesammelt, geordnet und aufbewahrt, und die Bezüge zu den Werken sind evident. Dagegen war die frühere Germanistik über die biographische „Aussagefähigkeit“ seiner Briefe eher enttäuscht. Vereinfacht könnte man sagen, dass sich Wedekind auch in seinen Briefen nicht mehr an die Konstruktionen der „privaten Briefe“ einer bürgerlichen Gefühlskultur hält. Dabei macht natürlich gerade dies einen hohen Quellenwert aus. Die zu ihrer Zeit als „persönliche“ Korrespondenzen entstandenen Texte in den Kreislauf der öffentlichen Kommunikation zurückzuführen, gehört auch zu den Zielen oder zumindest den Konsequenzen jeder elektronischen Briefedition.

Der Versuch, die semantischen Relationen der für Wedekinds Werk so wichtigen kulturhistorischen Kontexte abzubilden, erweist sich als einzigartige Chance wie als Risiko der Online-Edition.

Technische Konzeption

Im Folgenden sollen wichtige Aspekte der technischen Konzeption der Online-Briefedition beschrieben und die gewählten Lösungsansätze vorgestellt werden.

Eine der wichtigsten Zielstellungen der Online-Plattform ist die komplett webbasierte Umsetzung des Editionssystems. Dies bezieht sich zum einen auf die Präsentations- und Recherche-Ebene aber auch, und das ist unseres Wissens ein Alleinstellungsmerkmal für Online-Briefeditionen, auf die Erstellungs-Ebene. Auf diesen Aspekt wird später noch detaillierter eingegangen.

Um eine größtmögliche Unabhängigkeit von spezifischen Software-Produkten (beispielsweise dem verwendeten Datenbankmanagementsystem) zu erreichen und die Möglichkeit der Wiederverwendung und Integration einzelner Komponenten in anderen Projekten zu ermöglichen, wurde eine, aktuellen Software-Architektur-Standards entsprechende, Java EE Mehr-Schichten-Architektur unter Verwendung des Objekt-Relationalem-Mapping-Standards JPA entworfen und implementiert. Als Speicherformat in der Datenbank wurde die vom Konsortium der Text Encoding Initiative definierte XML-Repräsentationssprache TEI¹ gewählt, um später ggf. einen Datenaustausch bzw. eine Anbindung an andere Projekte zu ermöglichen.

Für den wissenschaftlichen Benutzer bzw. jeglichen interessierten Leser der Online-Briefedition werden unterschiedlichste Möglichkeiten der Recherche bereitgestellt: Neben der üblichen Volltextsuche wird die gezielte Suche nach Personen, Orten, Datumsangaben etc. angeboten. Darüber hinaus wird die gezielte Auswahl von Briefwechseln zwischen bestimmten Personen ermöglicht, um so Konversationsketten analysieren zu können. Für die benutzerfreundliche Gestaltung der Oberfläche wurden aktuelle Web-Technologien (JSF 2.0, PrimeFaces, CSS, AJAX etc.) verwendet.

Eine schon erwähnte Besonderheit dieses Projektes ist, dass auch die Erstellungs-Ebene, d.h. sowohl die Erfassung als auch die ggf. nachträgliche Bearbeitung komplett webbasiert realisiert wurde. Hintergrund dieser Entscheidung ist, dass für die Eingabe der transkribierten und kommentierten Daten mangels Personalressourcen mittelfristig nicht nur editionswissenschaftliche Experten und Literaturwissenschaftler zur Verfügung stehen, sondern diese Arbeit auch mit studentischen Hilfskräften ausgeführt werden soll. Daraus ergibt sich die zwingende Notwendigkeit, eine komfortable und fehlerreduzierende Eingabeoberfläche anzubieten. Während der Eingabe der Meta-Daten des Briefes (Verfasser, Empfänger, Datum des Briefs, Standort, Zustellweg etc.) wird der Benutzer unterstützt, in dem bereits erfasste Bezeichnungen – inklusive Alternativbezeichnungen – komfortabel zur Auswahl angeboten werden. Darüber hinaus ist in einem nächsten Schritt geplant, die semi-automatische Erkennung einer Named Entity (Namen, Orte etc.) im Rohtext des einzupflegenden Briefs zu unterstützen. Dazu sollen entsprechende Text-Mining-Verfahren in das Projekt integriert werden.

Eine besondere technische Herausforderung stellte die webbasierte Erfassung des Brief-Korpus mit allen editionswissenschaftlichen Auszeichnungen und Kommentierungen dar. Nach Prüfung der Integrations- bzw. Anpassungsmöglichkeiten verschiedener Web-Editoren haben wir uns aus technischen und lizenzrechtlichen Gründen für eine Eigenentwicklung unter Verwendung der Web-Technologien AJAX, jQuery und Javascript entschieden. Der entwickelte webbasierte TEI-Editor bietet die Möglichkeit, Auszeichnungen

¹ Text Encoding Initiative Consortium <http://www.tei-c.org>

und Kommentierungen komfortabel im gewohnten „Word-Feeling“ einzugeben. Abbildung 1 gibt einen kleinen Eindruck vom look and feel des Web-TEI-Editors.



Abbildung 1

Bei der Realisierung der Such- und Ausgabeoberfläche kommen aktuelle XML- und XSLT-Technologien zur Umwandlung in HTML bzw. PDF zum Einsatz. Die eigentliche Herausforderung liegt dabei allerdings in der Gestaltung des User Interfaces. Hier vollzieht sich ein grundlegender Medienwandel von der „Buchansicht“ einer gedruckten Ausgabe zu den Möglichkeiten einer webbasierten Darstellung. Eine bloße Transformation der „Buchansicht“ in das elektronische Medium greift hier zu kurz. Neben den Möglichkeiten einer parallelen Ansicht (beispielsweise von Transkriptionen und Original-Handschriften – siehe auch Abbildung 2) und den einschlägigen Navigationsmöglichkeiten zwischen den Briefen stellen sich auch grundlegende Fragen nach der angemessenen Visualisierung von Kommentaren, aber auch nach dem Umgang mit weiterführenden Informationen außerhalb der originären Briefeditionen.

The screenshot displays the homepage of the 'Editions- und Forschungsstelle Frank Wedekind' and its 'Online-Volltextdatenbank für Briefe von und an Frank Wedekind'. The header features a portrait of Frank Wedekind and the text 'Editions- und Forschungsstelle Frank Wedekind' and 'Online-Volltextdatenbank für Briefe von und an Frank Wedekind'. The main content area shows a letter to 'Geliebte Tilly'. An annotation box titled 'Einzelstellenkommentare' is overlaid on the text, containing the note: 'Gestern war ich mit Langheinrich,...' and 'Diesen Tag werde ich nie vergessen'. Below the letter, a sidebar displays a photograph of Frank Wedekind. The right side of the screen shows a detailed view of a handwritten letter from 'München' dated '20.11.14.', with some text visible in the background.

Abbildung 2

Fazit

Mit dem vorgestellten Projekt der Online-Brief-Datenbank wurde eine flexible Architektur und mächtige Software-Werkzeuge geschaffen, welche zum einen sowohl für den wissenschaftlichen Benutzer als auch andere interessierte Leser vielfältige und komfortable Möglichkeiten der Recherche bereitstellen. Zum anderen, und das ist der technische Hauptbeitrag dieses Projektes, ermöglichen sie eine komfortable und fehlerreduzierende Erfassung sowohl der Briefdaten als auch der Kommentierungen. Insgesamt wurde ein komplexes, aber eben auch einfach zu bedienendes Informationssystem für Online-Briefeditionen geschaffen.

Aus editionswissenschaftlicher Sicht ist entscheidend, dass die durch die Struktur der Datenbank deutlich gemachte Differenz zwischen Materialität, Befund und Deutung von Briefkorpora stets für die Wahrnehmung sowohl des Editors als auch des Nutzers klar und übersichtlich erhalten bleibt. Nichts wäre unfruchtbare, als im Dschungel einer überkomplexen Datenbank sich zu verirren. Um es noch einmal anders zu formulieren, nicht entscheidend ist das Prinzip, dass ‚alles‘ machbar sei, sondern ‚wie‘ ‚was‘ gestaltet ist. Das sollte stets von Anfang an Ziel eines Datenbank-Modells für Editionen sein. Daraus entspringt im eigentlichen Sinn die Usability einer online-Briefdatenbank. Dagegen sollte der ‚Verwertung‘ des zu Schaffenden und des Geschaffenen keine Grenzen gezogen werden. Gemeint ist damit deren beider Zugänglichkeit.

APW digital – Perspektiven einer digitalisierten Edition

Vorschlag eingereicht von Tobias Tenhaef M.A., Wissenschaftlicher Mitarbeiter des DFG-geförderten Digitalisierungsprojekts „APW digital“ (ttenhaef@uni-bonn.de), und von Dr. Dr. Guido Braun, operativer Projektleiter (gbraun@uni-bonn.de)

Die ACTA PACIS WESTPHALICAE (APW) sind die historisch-kritische Edition der wichtigsten Quellen des Westfälischen Friedenskongresses. Bis Herbst 2013 erschienen 48 Bände mit den zentralen Akten der Verhandlungen zu den Friedensschlüssen von Kaiser und Reichsständen mit Frankreich und Schweden vom 24. Oktober 1648. In Sachkommentaren und Einleitungen wird in diesen Bänden auch der Verhandlungsgang erschlossen, der zu dem spanisch-niederländischen Friedensvertrag vom 30. Januar 1648 führte. Dasselbe gilt für die spanisch-französischen Verhandlungen, die in Westfalen ohne Ergebnis blieben. Die drei umfangreichen Verträge des Jahres 1648 schufen in Deutschland und in einem Teil von Europa dauerhafte Befriedung. Für das Reich entstanden funktionierende Regeln für das friedliche Zusammenleben in einem mehrkonfessionellen Gemeinwesen, die noch im heutigen deutschen Staatskirchenrecht nachwirken. Von Europa aus gesehen war der Kongress in Münster und Osnabrück die erste große, nicht kirchlich geprägte Versammlung, die die Entwicklung des europäischen Mächtesystems erheblich beförderte. Darüber hinaus besitzt der Westfälische Frieden eine paradigmatische Bedeutung für das Friedenschließen überhaupt.

Die Edition APW wurde von 1957 bis 2011 von den wissenschaftlichen Mitarbeiterinnen und Mitarbeitern der "Vereinigung zur Erforschung der neueren Geschichte" angefertigt, seit 1977 als Projekt der Union der Akademien. 2013 ging die Arbeitsstelle Westfälischer Frieden 1648, die von der „Vereinigung“ in Bonn unterhalten wurde, im neu gegründeten Zentrum für Historische Friedensforschung der Philosophischen Fakultät der Universität Bonn auf, das vom Lehrstuhlinhaber für die Geschichte der Frühen Neuzeit, Maximilian Lanzinner, geleitet wird.

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG), wurden bis Ende 2012 etwa 28.000 Seiten der bis 2008 erschienenen 40 Editions-Bände retrodigitalisiert und um digitale Zusatzangebote erweitert. Ziel des Projekts APW digital ist es, die bisher nur in gedruckter Form vorliegende Edition der APW als digitalen Volltext im World Wide Web zur Verfügung zu stellen. Getragen wird das Projekt vom Bonner Lehrstuhl für die Geschichte der Frühen Neuzeit in Verbindung mit dem Zentrum für Historische Friedensforschung und von der Bayerischen Staatsbibliothek München (BSB). Derzeit werden die von der Bonner Projektgruppe bereitgestellten Daten bei der BSB aufbereitet. Die Online-Stellung ist für das erste Quartal 2014 vorgesehen.

Das Digitalisierungsprojekt "APW digital" hat eine Pilotfunktion im Bereich der Online-Editionen zur frühneuzeitlichen Geschichte. Vor den APW waren international nur wenige Editionen zur Frühen Neuzeit im Internet abrufbar, die jedoch entweder nur einen kleineren Quellenbestand bereitstellten oder bei ähnlich umfassender Quellenaufbereitung nicht die gleichen Arbeitsmöglichkeiten anboten wie "APW digital".

Diese Möglichkeiten der digitalisierten Edition übertreffen jene der Print-Fassung bei Weitem. Unterschiedliche Suchfunktionen erlauben durch verschiedene Recherchemöglichkeiten eine verbesserte Erschließung und damit eine tiefere wissenschaftliche Nutzung der Texte.

APW digital bietet **vier Zugriffsmöglichkeiten** auf die Texte der Edition:

1. den Zugriff über die **Struktur der gedruckten Edition** nach Serien, Abteilungen und Bänden; die Einteilung der Printfassung ist erhalten;
2. den streng **chronologischen Zugriff** auf die edierten Dokumente,
3. den **Direktzugriff** auf einzelne Dokumente und Seiten,
4. die Möglichkeit einer unscharfen **Volltextsuche**, wie er nur bei digitalen Volltexten gegeben ist.

Die Volltextsuche wird ergänzt durch die Register, die bereits in den Printbänden als Erschließungs- und Recherchesysteme vorhanden waren. Sie wurden vollständig in die APW digital übernommen. Mit der Umsetzung der bändeübergreifenden Verweise als Hyperlinks wird das schon in der Druckfassung angelegte Potential des Hypertextes digital realisiert. Die Verknüpfung mit anderen Digitalisaten im Netz erleichtert die Benutzbarkeit gegenüber der Printfassung der APW deutlich. Abgerundet wird das Angebot von APW digital durch die Bereitstellung von wertvollen Verknüpfungen, die teils nur im Netz möglich sind: zu den Personen, die am Westfälischen Friedenskongress teilgenommen hatten und im Rahmen des Projekts mit GND-Nummern eindeutig identifiziert wurden; zu den genannten Orten, die georeferenziert wurden und als Basis einer Kartendarstellung dienen; schließlich zu den Ereignissen selbst mithilfe zweier chronologischer Übersichten ("für Einsteiger" bzw. "für Experten").

"APW digital" ist anschlussfähig für die Integration weiterer Editionsbände, die teils schon vorliegen, teils in Bearbeitung sind. Ziel des geplanten Vortrages ist daher nicht allein die Vorstellung der im März 2014 voraussichtlich online einsehbaren digitalen Edition der ersten 40 APW-Bände, die Gegenstand der DFG-Förderung waren.

Auf der Basis des bestehenden Online-Angebots "APW digital" werden vielmehr weitergehende Perspektiven eröffnet, die eine mögliche zukünftige Fortentwicklung des Projektes leiten können. Dabei wird insbesondere zu diskutieren sein, inwiefern das bei APW digital zur Verfügung stehende maschinenlesbare Korpus für linguistische und sprachgeschichtliche Fragestellungen genutzt werden kann und welche Erweiterungen mittel- und langfristig wünschbar sind. Aber auch die Frage, wie APW digital zur GIS-gestützten historischen Geowissenschaft und Kartographie beitragen kann und welche Bedingungen erfüllt sein müssen, um APW digital zu einem Teil des Semantischen Netzes werden zu lassen, sollen eingehend diskutiert werden.

Die Erörterung dieser Perspektiven soll im Mittelpunkt des geplanten Vortrages stehen (ca. zwei Drittel der vorgesehenen Vortragszeit).

Manfred Thaller, Universität zu Köln und Walter Scholger, Universität Graz

Panel: Digital Humanities als Beruf - Der Weg zu einem Curriculum

Dass die Digital Humanities derzeit eine echte Boomphase durchleben, ist unbestreitbar. Die Anzahl nationaler und internationaler Projekte dazu ist in den letzten Jahren sprunghaft angestiegen. Dabei ist die Situation der deutschsprachigen Länder ungewöhnlich dadurch, dass die Anzahl der hier als durchstrukturierte Studiengänge angebotenen Abschlüsse – zum Unterschied von kursartig angebotenen Zusatzqualifikationen – deutlich über denen anderer Länder liegt. Dem entspricht, dass seit dem November 2009 eine lose Gruppe von VertreterInnen dieser Studiengänge sich mehrfach an der Universität zu Köln getroffen hat, um einen Gedankenaustausch über Gemeinsamkeiten und Unterschiede dieser Studiengänge einzuleiten. Dabei war zu hoffen, dass sich daraus Möglichkeiten ergeben, aus den Gemeinsamkeiten dieser Studiengänge Eckwerte abzuleiten, die letzten Endes zu einem Referenzcurriculum zusammengefasst werden könnten. Begonnen zunächst als eine lokale Initiative, wurde dieser Diskussionsprozess später von Dariah-DE aufgegriffen. Im Rahmen dieses Infrastrukturprojekts wurde 2011 eine Übersicht über einschlägige deutsche Studiengänge veröffentlicht¹, der Ende 2013 als weiteres Zwischenergebnis ein Versuch der Kategorisierung bestehender Studienangebote folgte².

Wozu ein „Referenzcurriculum“, wenn die Lehrangebote sich offensichtlich auch ohne ein solches dynamisch entwickelt haben? Unstrittig wurden in den letzten Jahren, national wie international, durch Fortbildungen in Form von Summer Schools, Thatcamps, Workshops und Einzelkursen im Hintergrund enorme Anstrengungen zur Entwicklung von Lehrangeboten unternommen., zu denen, wie einleitend angemerkt, gerade im deutschen Sprachraum zunehmend auch voll etablierte Studiengänge treten, wobei die diversen Bachelor, Master und Promotionsangebote die individuelle Situation der einzelnen Hochschulen und Institutionen wiederspiegeln. Wir finden jedoch, dass diese Konjunktur nicht dazu verführen sollte, zu glauben, dass diese Ausweitung des Angebotes bereits selbsttragend sei. Man darf entsprechende Erfahrungen der Vergangenheit nicht außer Acht lassen. Im damaligen Status-Report *Computing in Humanities Education: A European Perspective*³ von 1999 wurden 25 verschiedene Digital Humanities Studiengänge vorgestellt und diskutiert – gerade einmal neun existieren davon noch, von denen wiederum fünf verschiedene Auslegungen der Computer Linguistik repräsentieren, bei denen die Zuordnung zu den Digital Humanities nicht unbedingt eindeutig ist. In den frühen Neunzigern wurden auf zwei Workshops zum Zwecke der Entwicklung eines internationalen Curriculums für *History and Computing* etwa 15 Studiengänge an europäischen Einrichtungen präsentiert, von denen heute noch genau zwei existieren – einer unter massiver thematischer Neuorientierung. Von sechs italienischen Angeboten aus dem gleichen Zeitraum ist exakt ein einziges verbliebenen.

Die jetzige Konjunktur der interdisziplinären Arbeit zwischen den Geisteswissenschaften und der Informatik sollte daher nicht in einer Bestandsaufnahme nach dem Modell von 1999 stecken bleiben, sondern versuchen über das Stadium „Digital Humanities Kurse unterrichten, was die an den jeweiligen Universitäten Digital Humanities Unterrichtenden unterrichten“ hinaus zu kommen und von persönlichen Forschungsrichtungen und lokalen Gegebenheiten zu abstrahieren. Dies ist nicht einfacher, als einer der aktuellen Versuche, die Digital Humanities additiv als solche zu definieren⁴. Es ist aber notwendig, aus pragmatischen Gründen:

- Je größer die Zahl einschlägiger Studiengänge wird, desto schwieriger ist es zu vermitteln, warum der Übergang von einem zum anderen Probleme bereiten sollte. Die wechselseitige Anerkennung von Studienleistungen wird erheblich vereinfacht, wenn sie studiengangsunabhängig definiert sind.
- Die Akkreditierung von Studiengängen wird umso einfacher, je einfacher es ist, sich bei dem Studiengang auf einverständlich über einzelne Institutionen hinaus definierte Referenzwerte zu beziehen.
- Definieren die DH Studiengänge ihre eigenen Orientierungspunkte nicht selbst, ist durchaus zu erwarten, dass andere versuchen, dies für sie zu tun.

¹ <http://www.cceh.uni-koeln.de/Dokumente/BroschuereWeb.pdf>

² Patrick Sahle: "[DH Studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities](#)". Dariah-DE Working Papers Nr. 1. Göttingen: Dariah-DE, 2013. URN: urn:nbn:de:gbv:7-dariah-2013-1-5.

³ <http://www.hd.uib.no/AcoHum/book/>

⁴ Zuletzt Melissa Terras, Julianne Nyhan, Edward Vanhoutte (Edd): Defining Digital Humanities: A Reader, Ashgate, 2013.

Dabei kann – und darf – es nicht darum gehen, in einem sich nach wie vor sehr dynamisch weiter entwickelnden Bereich verbindliche Details, etwa im Sinne einer verpflichtenden Studienordnung, festzuschreiben: Der Begriff eines „Referenzcurriculums“ versteht sich bewusst im Sinne einer Referenzarchitektur, nach dem Gebrauch des Begriffs in der Softwaretechnologie. Es soll also einerseits ein Modell beschreiben, mit dem einzelne konkrete Curricula verglichen werden können, andererseits ein Vokabular definieren, mit dessen Hilfe Umsetzungen möglichst präzise definiert werden können.

Die Bemühungen der curricularen Arbeitsgruppe manifestieren sich auch in einer engen Kooperation mit der Arbeitsgruppe „Training and Education“ in Dariah-EU. Die Entwicklung eines europäischen Referenzrahmens für die Kompatibilität nationaler Bildungsangebote im Bereich der Digital Humanities, sowohl hinsichtlich inhaltlicher Bausteine als auch curricularer Anrechenbarkeiten, ist ein zentrales Anliegen dieser europäischen Infrastrukturinitiative. KollegInnen aus Deutschland, Österreich und der Schweiz beteiligen sich daher im Kontext ihrer Beiträge zu Dariah-EU unmittelbar an den Diskussionen und bringen die Ergebnisse in den breiteren europäischen Diskurs ein.

Der erreichte Stand dieser Überlegungen wird in Passau präsentiert werden und eine Gruppe der an seiner Vorbereitung beteiligten Kolleginnen und Kollegen wird in persönlichen Statements einzelne Positionen dazu vertreten und diskutieren, bevor die Diskussion für das Publikum geöffnet wird. Dabei gehen wir von einem Zeitverhältnis Präsentation : Paneldiskussion : Publikumsdiskussion von 1 : 1 : 1 aus.

Die sechs TeilnehmerInnen am Panel sind noch nicht abschließend bestimmt. Die curriculare Arbeitsgruppe besteht derzeit aus folgenden Damen und Herren:

Sabine Bartsch, Technische Universität Darmstadt; Michael Beisswenger, Technische Universität Dortmund; Frank Binder, Universität Gießen; Stefan Büttner, Fachhochschule Potsdam; Elisabeth Burr, Universität Leipzig; Marcus Held, Institut für Europäische Geschichte, Mainz; Andreas Henrich, Universität Bamberg; Ansgar Kellner, Universität Göttingen; Matthias Lang, Universität Tübingen; Andy Lücking, Universität Frankfurt; José Manuel Martínez Martínez, Universität des Saarlandes; Klaus Meyer-Wegener, Universität Erlangen-Nürnberg; Matthias Perstling, Universität Graz; Steffen Pielström, Universität Würzburg; Malte Rehbein, Universität Passau; Patrick Sahle, Universität zu Köln; Andrea Schneider, Göttingen Center for Digital Humanities; Markus Schnoepf, Berlin-Brandenburger Akademie der Wissenschaften; Christof Schöch, Universität Würzburg; Walter Scholger, Universität Graz; Caroline Sporleder, Universität Trier; Maik Stührenberg, Universität Bielefeld; Manfred Thaller, Universität zu Köln; Armin Volkmann, Universität Heidelberg;

Vortrag 1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014),
Universität Passau, 25.-28. März 2014

Entwicklung des digitalen Tanzarchivs der Pina Bausch-Stiftung

Bernhard Thull, Vera Marz

Motivation und Ausgangspunkt

Das Werk der international anerkannten Choreografin Pina Bausch umfasst mehr als 50 Stücke, die weltweit aufgeführt worden sind. Nach ihrem Tod im Jahre 2009 hat es sich die Pina Bausch-Stiftung zur Aufgabe gemacht, dieses Werk zu bewahren. Es ist durch umfangreiches Material z.B. im Archiv der Pina Bausch Stiftung dokumentiert. Vielleicht noch wichtiger ist aber die Erinnerung von vielen Menschen in der ganzen Welt an ihr Werk. Das Werk von Pina Bausch zu bewahren bedeutet daher sowohl Material als auch Erinnerungen zu erfassen und zu erhalten.

Das Material im Archiv der Pina Bausch Stiftung umfasst Stück- und Aufführungslisten, Tonaufnahmen und ihre Transkriptionen, Regiebücher, Dokumentationen der Bühnenbilder, Fotografien und Videoaufnahmen, Kostüme, Manuskripte, Requisiten, Tänzeraufschriebe und vieles mehr. Sie enthalten Informationen und inhaltliche Beziehungen z.B. über Personen, Stücke und ihre Aufführungen, Besetzungen oder Rollen und ihre Entwicklung. Obwohl der größte Teil des Materials im Archiv der Pina Bausch Stiftung eingelagert ist, ist weiteres Material in Tanzarchiven weltweit verteilt oder nicht einmal Teil eines Tanzarchivs, wie z.B. ein Zeitungsartikel über eine ganz bestimmte Aufführung in einem Zeitungsarchiv oder ein Buch über das Werk von Pina Bausch in einer Bibliothek.

Erinnerungen sind mit Personen, Aufführungen, Szenen, oder vielleicht mit einem Bühnenbild oder mit einer bestimmten Musik verbunden und sie beziehen sich aufeinander und auf das Material in individueller Weise. Sie sind unter den Mitgliedern des Tanztheater Wuppertal Pina Bausch, aber auch unter vielen anderen Menschen weltweit verteilt. Im Vergleich zu dem gesammelten physischen Material des Archivs sind menschliche Erinnerungen von Natur aus emotional, ungenau, widersprüchlich, inkonsistent und unvollständig. Sie liefern Fragmente, die zusammen betrachtet ein Bild ergeben.

Mit dem digitalen Archiv verfolgt die Stiftung verschiedene Ziele. Das Archiv soll z.B. dem Tanztheater Wuppertal Pina Bausch dabei helfen, Stücke wieder aufzuführen, Tanzinteressierten das Werk von Pina Bausch vermitteln oder Wissenschaftlern als Werkzeug für ihre Forschung dienen.

Anforderungen

Welche Werkzeuge unterstützen die Sammlung von Material und von Erinnerungsfragmenten und erlauben dann, daraus größere Bilder entstehen zu lassen? Die wesentlichen Anforderungen im Detail:

- Beschreibung des physischen Materials, das von Pina Bausch gesammelt worden ist. Dieses Material sollte erfasst und Beziehungen zwischen diesem Material repräsentiert werden können.
- Sammlung von Erinnerungen sowie jegliche gesprochenen oder aufgeschriebenen Anmerkungen über jede Art von Material des Archivs. Beispielsweise sollte eine Tänzerin in der Lage sein, eine bestimmte Szene eines Stücks mündlich zu kommentieren oder ein Zuschauer sollte seine Erfahrungen mit einer bestimmten Aufführung beschreiben können.
- Beliebiges Material miteinander verbinden. Beispielsweise sollte es möglich sein, eine bestimmte Requisite mit einem Manuscript zu verbinden und festzuhalten, dass die Gestaltung dieser Requisite in diesem Manuscript beschrieben worden ist. Oder ein Zuschauer hat einen seltenen Zeitungsartikel aus den siebziger Jahren gefunden und möchte ihn mit der Aufführung verbinden, von der dieser Artikel berichtet.
- Aggregation, Sortierung, Klassifikation oder eine andere Art der Verarbeitung des Materials, um es für die Entwicklung von Interpretationen, Visualisierungen oder interaktive Erfahrungen aufzubereiten. Dies könnten Untersuchungen im Rahmen von Forschungsarbeiten sein, Websites für ein bestimmtes Publikum wie z.B. Kinder, oder interaktive Installationen im Rahmen von Ausstellungen.

Ansatz *Linked Data*

Die zu erfüllenden Anforderungen erinnern an die Situation im *World Wide Web*. Das Web ist ein verteiltes System von miteinander verbundenen Websites. Einige Teile des Web sind aufgeräumt, gut strukturiert und organisiert, andere Teile eher spontan und chaotisch. Jeder kann Inhalte beisteuern, die letztendlich ein einziges Netz von Dokumenten (*web of documents*) bilden. Das World Wide Web Consortium¹ (W3C) hat erkannt, dass die weiter oben genannten Anforderungen sehr verbreitet sind und für viele unterschiedliche Anwendungsbereiche gelten. Es hat daher eine neue Form der Datenmodellierung vorgeschlagen, die *Linked Data*² heißt, und die es erlaubt, das *web of documents* in ein *web of data* umzuwandeln, ohne die offene und dynamische Natur des Web aufzugeben. Die Ergänzung des *Linked Data*-Paradigmas mit Werkzeugen, die automatisch logisch konsistente Datensätze über verteilte Daten aufbauen können, erlaubt die Entwicklung des so genannten *Semantic Web*³. Es erscheint sinnvoll zu untersuchen, inwieweit sich das *Linked Data*-Paradigma als Ansatz für das digitale Archiv der Pina Bausch Stiftung eignet.

¹ <http://www.w3.org>

² <http://www.w3.org/standards/semanticweb/data>

³ <http://www.w3.org/standards/semanticweb>

Stand des Archivs

Experimenteller Aufbau

Um zu untersuchen, ob der *Linked Data*-Ansatz als Basis für das digitale Archiv der Pina Bausch Stiftung geeignet ist, haben wir eine experimentelle Systemarchitektur entwickelt. Daten werden aus vielen verschiedenen Quellen gesammelt, wie z.B. aus einer vorhandenen Filemaker *Bento*-Datenbank mit Daten über Kostüme, Microsoft *Excel*-Tabellen oder einfach *Linked Data*-Dateien, die in RDF⁴ oder Turtle⁵ geschrieben sind und z.B. die Beschreibung der Ontologie enthalten. Der *Triple Store* selbst ist mit Hilfe des OWLIM lite⁶ *Triple stores* in Verbindung mit der OpenRDF Sesame Workbench⁷ implementiert. Mit Hilfe eines einfachen web-basierten Datenbrowsers, der im Rahmen des Projekts entwickelt worden ist, kann man Daten des Archivs einsehen, Daten in das Archiv einlesen und aus dem Archiv löschen. Er dient als Werkzeug, um die Korrektheit der Daten und ihrer Verlinkung mit anderen Daten zu überprüfen.

Prozesse und Daten im Archiv

Da der *Linked Data*-Ansatz kein a priori-Datenmodell benötigt, um Daten zu erfassen, kann man die Prozesse der Datenmodellierung und der Datenerfassung trennen und den Entwicklungsprozess des Archivs beschleunigen. Der Datenerfassungs- und der Datenmodellierungsprozess laufen daher unabhängig voneinander. Die Eingabe von Massendaten über das Material des Archivs erfolgt zurzeit mit Hilfe von *Excel*-Tabellen. Bis jetzt wurden insgesamt fast 48.000 Entitäten erfasst und mit knapp 920.000 Tripel beschrieben, darunter:

- 466 Personen
- 54 Stücke mit insgesamt 3.081 Szenen
- 6.286 Aufführungen
- 12.951 Objekte (z.B. Programmhefte, Videos, Fotografien, Poster oder Dokumente)

Die *Excel*-Tabellen sind dabei so angelegt, dass sie keine Information über die Modellierung enthalten. Die Modellierung der Daten erfolgt erst beim Einlesen der Tabellen in die Datenbank. Dieses Vorgehen erlaubt es, das Modell jederzeit zu überarbeiten, ohne bereits erfasste Daten wieder zu verlieren.

⁴ http://www.w3.org/standards/techs/rdf#w3c_all

⁵ <http://www.w3.org/TR/turtle/>

⁶ <http://www.ontotext.com/owlim>

⁷ <http://www.openrdf.org/>

Aktuelles Modell

Die zunächst wichtige Archivperspektive haben wir mit Hilfe des Modells der *Functional Requirements of Bibliographic Records* (FRBR, [1]) realisiert. Das FRBR-Modell dient als Ausgangspunkt für die Entwicklung der *Pina Bausch Archive Ontology*, das wir nach Bedarf instanziieren und ergänzen. Dabei verwenden wir etablierte Vokabulare, wie z.B. *Dublin Core Metadata Terms*⁸ oder *Simple Knowledge Organization System*⁹, um die Einhaltung von *Linked Data*-Prinzipien sicherzustellen.

Fazit

Wir haben genügend Daten erfasst, um die Machbarkeit und die Sinnhaftigkeit des *Linked Data*-Ansatzes für das digitale Archiv der Pina Bausch Stiftung zu zeigen. Insbesondere das FRBR-Modell konnte sein Versprechen halten und hat sich bewährt. Auch wenn die verwendete Technologie des *Triple Stores* und der dazugehörigen Werkzeuge noch relativ jung im Vergleich zu relationalen Datenbanken ist, zeigte sie sich ausreichend gereift und stabil. Dieser experimentelle Aufbau war daher ein erster wertvoller Schritt zur Entwicklung des digitalen Archivs, das durch den Einsatz eines Entwicklungswerkzeuges wie z.B. *Callimachus*¹⁰ weiter professionalisiert werden kann.

Die Benutzungsschnittstelle und die Visualisierung sowohl zur Eingabe von Daten als auch zur Erkundung von Daten hat sich dagegen als eine wesentliche Herausforderung herausgestellt. Der *Linked Data*-Ansatz widersetzt sich üblichen Methoden des *User Interface Designs*, wie wir im Projekt an vielen Stellen gelernt haben. Dies hängt mit der Dynamik von *Linked Data*-Archiven zusammen, wo eine Ressource mit nur ein paar andere Ressourcen verbunden sein kann, wohingegen eine andere Ressource der gleichen Art mit Hunderten anderen Ressourcen verbunden sein kann, wie z.B. Stücke und ihre Aufführungen. Weitere Arbeiten werden einen starken Fokus auf diesen Teil des digitalen Archivs legen müssen.

Zusammenfassend lässt sich feststellen, dass der *Linked Data*-Ansatz die Anforderungen an das digitale Archiv der Pina Bausch Stiftung zu erfüllen scheint. Durch den Einsatz von *Linked Data* und die Fähigkeit des *Triple Store*, fehlende Verbindungen herzustellen, ist es möglich, lokales und verteiltes Wissen zu einer Sicht zusammenzuführen. Mit Hilfe von Vokabularen und Ontologien, die zwischen verschiedenen Vokabularen vermitteln können, ist es möglich, verschiedenen Sichten auf und Interpretationen des Materials zu entwickeln. Das Konzept so genannter Kontexte erlaubt es, Autorschaft nachzuhalten und somit Widersprüche selbst auf der Ebene von Fakten darzustellen. Die Möglichkeit, Daten ohne a priori-Modellierung zu erfassen, erlaubt es, sich dynamisch an veränderte Anforderungen

⁸ <http://dublincore.org/documents/dcmi-terms/>

⁹ http://www.w3.org/standards/techs/skos#w3c_all

¹⁰ <http://callimachusproject.org>

und Randbedingungen anzupassen. Und als ein letzter, wichtiger Punkt stellt die Einhaltung der *Linked Data*-Prinzipien¹¹ sicher, dass das Archiv zukünftig leicht mit anderen *Linked Data*-Archiven verbunden werden kann.

- [1] IFLA Study Group on the Functional Requirements for Bibliographic Records:
Functional requirements for bibliographic records, Final report, International Federation of Library Associations and Institutions, 2009

¹¹ <http://www.w3.org/DesignIssues/LinkedData.html>

Schrift und Zeichen. Computergestützte Analyse von hochmittelalterlichen Papsturkunden. Ein Schlüssel zur Kulturgeschichte Europas

Das Projekt *Schrift und Zeichen. Computergestützte Analyse von hochmittelalterlichen Papsturkunden. Ein Schlüssel zur Kulturgeschichte Europas* wird seit Juni 2012 vom BMBF im Rahmen der eHumanities gefördert. Die drei Teilprojekte Paläographie, Mittelalterliche Geschichte und Informatik beschäftigten sich mit der computergestützten Erfassung, Analyse und Kategorisierung der Schrift und der Layoutmerkmale hochmittelalterlicher Papsturkunden, einem der wichtigsten und umfangreichsten Quellenkorpora.

So zählte die päpstliche Kurie neben der Kanzlei der römisch-deutschen Könige und Kaiser zu den bedeutendsten Urkundenausstellern des Mittelalters und dies nicht nur im Hinblick auf die Quantität, sondern auch der Qualität der Urkunden. Für den Zeitraum von 753 bis 1198 sind rund 25.000 Papsturkunden überliefert, die von der römischen Kurie aus an die gesamte christliche Welt gingen. Dort fand nicht nur der Rechtsinhalt Beachtung, sondern auch die Form der Urkunden, da diese vielfach als Vorbild für die lokale Urkundenproduktion dienten. Jedoch können Papsturkunden nicht als starres Formular betrachtet werden, welches einmal eine Form angenommen sich nicht mehr verändert hat. Im Gegenteil: Die äußere Form und vor allem die verwendete Schrift veränderten sich im Verlauf des Untersuchungszeitraumes von 1054 bis 1198 auf vielfältige Weise. So wechselte die Urkundenschrift von der päpstlichen Kurialen zur karolingischen Minuskel und schließlich zur gotischen Urkundenschrift. Dem Ansatz von Heinrich Fichtenau folgend, möchte das Projekt die Schrift nicht als bloßen Informationsträger verstehen, sondern als ein Kulturgut, an dessen Umgestaltung sich kulturelle Veränderungen widerspiegeln. Daher stehen im Fokus des Projektes Fragen nach dem Verhältnis von Schriftveränderung und graphischen Symbolen, Abhängigkeit der Schriftvarianz von Empfängern, Inhalten oder einzelnen Schreiberhänden, sowie die Eigenhändigkeit der Unterschriften der Päpste und Kardinäle.

Für die Klärung dieser Fragen muss sich das Projekt vor allem zwei Hauptthemen widmen: der Beschreibung der Schriftveränderung und einer Schreiberidentifizierung.

Die detaillierte Beschreibung der Schriftveränderung erfordert eine enge Zusammenarbeit der drei Teilprojekte. So wird zunächst von dem Teilprojekt Paläographie ein Merkmalskatalog zu den verschiedenen in den Papsturkunden verwendeten Schriften entwickelt. Dabei werden die Buchstaben bis in Einzelemente untergliedert, um Schriftspezifika besser beschreiben zu können. Auf Grundlage dieses Merkmalskatalogs entwickelt das Teilprojekt Informatik Tools, welche die Schriftentwicklung nachvollziehbar machen sollen. Für die Analyse der Buchstabenformen und deren zeitliche Entwicklung werden Methoden der Mustererkennung

benutzt. Zum Erlernen der verschiedenen Formen wird das System zunächst trainiert, indem die einzelnen Symbole und Buchstaben ausgezeichnet werden („supervised learning“). Das hierfür entwickelte Annotationswerkzeug ermöglicht eine Kommentierung der jeweiligen Zeichen, die in XML-Strukturen zur weiteren Verwendung gespeichert werden. Schließlich sollen für die eigentliche Analyse der Buchstaben verschiedene Klassifikatoren getestet und evaluiert werden. Als Merkmale dienen unter anderem der Neigungswinkel, die Strichstärke, die Länge der Schäfte im Verhältnis zum Körper des Buchstabens oder ob ein Zeichen einen Bogen oder einen Knick enthält. Eine Auswertung dieser Veränderung der Schrift erfolgt dann im Teilprojekt Geschichte. Dabei können durch das automatisierte Verfahren besonders hohe Datenmengen verarbeitet und auf diese Weise übergreifende Fragestellungen beantwortet werden, was sich gerade in Bezug auf das Kanzleiwesen als fruchtbringend erweist.

Darüber hinaus wird an einer Schreiberidentifizierung gearbeitet. Aufgabe der Paläographie dabei ist es, möglichst signifikante Merkmale einer Schreiberhand durch die Methode des Vergleichs zu identifizieren. Dazu eignet sich im besonderen Maß die Datumszeile, da dieser Teil der Urkunde mit einem (angeblichen) Schreibernamen versehen ist. Die Informatik versucht die Umsetzung dieses Vorhabens mit Hilfe der Betrachtung einzelner Worte, da erst in der Abfolge von mehreren Buchstaben schreiberspezifische Eigenheiten besonders hervortreten. Auf Grundlage von Algorithmen der Mustererkennung lassen sich dann Wahrscheinlichkeiten errechnen, ob es sich bei ausgewählten Schriftproben um die gleiche Hand handelt. Die Auswertung der Ergebnisse bietet Aufschluss über die Kanzlei in Bezug auf ihre personelle Zusammensetzung sowie deren Institutionalisierungsprozess. Des Weiteren werden dadurch auch tiefere Einblicke in das Kardinalat ermöglicht, etwa durch den Nachweis der Eigenhändigkeit der Kardinalsunterschriften auf Privilegien. Gerade in dieser frühen Phase des Kardinalskollegiums kann hier ein wichtiger Beitrag zur Bedeutung des Kardinalsranges gegeben werden.

Die hier skizzierten Einblicke in das Forschungsvorhaben des Projektes *Schrift und Zeichen* zeigen, neben den projektspezifischen Herangehensweise auch die praktische Umsetzung der Zusammenarbeit von Geisteswissenschaften und Informatik. So ist der ständige Austausch zwischen den Disziplinen ein unverzichtbarer Projektbaustein, der zu einer Horizonterweiterung auf beiden Seiten führt. Die Informatik ist nicht bloßer Dienstleister der Geisteswissenschaften, sie kann sich viel mehr in der Zusammenarbeit neue Forschungsfelder erschließen und die Geisteswissenschaften werden nicht etwa durch automatisierte Verfahren

ersetzt, sondern erschließen sich durch die automatische Aufarbeitung großer Quellenkorpora neue Forschungsfelder.

CLARIN-D und Forschungsinfrastrukturen

Thorsten Trippel, Dieter van Uytvanck

In den Geistes- und Sozialwissenschaften wird im Kontext der Digital Humanities von Forschungsinfrastrukturen gesprochen. Anders als in ingenieurwissenschaftlichen Bereichen oder in naturwissenschaftlichen Labors geht es bei diesen Einrichtungen nicht um die Aufstellung technischer Geräte, sondern um ein komplexes Zusammenspiel von Forschungsdaten, Archiven und Programmen zur Analyse der Daten. Im Bereich der europäischen Forschung im Rahmen des European Strategy Forum on Research Infrastructures (ESFRI) werden diejenigen Einrichtungen als Forschungsinfrastrukturen bezeichnet, die Ressourcen oder Dienste anbieten, die von Wissenschaftlern zur Durchführung Ihrer Forschung verwendet werden (siehe etwa die ESFRI-Broschüre der Europäischen Kommission, S. 4 [1]). Bereits bei der Vorstellung der ESFRI-Roadmap 2006 wurde die Funktion von Forschungsinfrastrukturen folgendermaßen beschrieben:

This definition of Research Infrastructures, including the associated human resources, covers major equipment or sets of instruments, as well as knowledge-containing resources such as collections, archives and databases. Research Infrastructures may be “single-sited”, “distributed”, or “virtual” (the service being provided electronically). They often require structured information systems related to data management, enabling information and communication. These include technology-based infrastructures such as grid, computing, software and middleware.

(ESFRI Roadmap, 2006, S. 14 [2])

Sprachbasierte Ressourcen, seien es Text-, gesprochene Sprache oder multimodale Daten stellen besondere Anforderungen an eine Infrastruktur. Diese kann nicht nur ein spezialisiertes Archiv für Forschungsdaten sein, sondern sie muss die speziellen Anforderungen, insbesondere rechtliche und ethische Richtlinien, berücksichtigen. Daher ist die CLARIN Infrastruktur selbst Ort und Werkzeug, um Forschung zu ermöglichen und komplexe Suchfunktionen und weitere Funktionen zur Inhaltserschließung anzubieten und Analysen auf den Ressourcen vorzunehmen.

CLARIN-D ist das nationale Projekt, das zur europäischen Forschungsinfrastruktur CLARIN gehört, die als European Research Infrastructure Consortium (ERIC) organisiert ist. Nachdem bislang primär die *Implementation* der grundlegenden Infrastrukturelemente Vordergrund stand und diese in Teilen bereits verfügbar sind, z.B. Repositorien, Vernetzung der Suchfunktionen über Repositorienkataloge hinweg, Dienste zur Erschließung von Ressourcen über Webservices usw., liegt nun der Fokus auf der *Anwendung* in den Geistes- und Sozialwissenschaften.

Für Forschungsinfrastrukturen im Bereich der eHumanities sind Datenrepositorien als Grundlage für wissenschaftliche Untersuchungen unerlässlich. Sie müssen ergänzt werden durch fachspezifische Werkzeuge zur Suche, Analyse und Visualisierung der Daten. In seinem Aufsatz „Controversies around the Digital Humanities: An Agenda“ [3] weist Thaller darauf hin, dass die Möglichkeiten der Analyse bisher mit der Zunahme verfügbarer Ressourcen nicht Schritt halten. Die Analyse innerhalb der Infrastruktur von CLARIN begegnet dieser Kritik durch eine skalierbare Services und durch einen Mehrwert durch die Kombination der verteilt eingesetzten Werkzeuge und Ressourcen. CLARIN geht auf die Geistes- und Sozialwissenschaften zu, die sprachbasierte Daten verwenden, um passende Werkzeuge anzubieten und gegebenenfalls anzupassen.

In dem Vortrag stellen wir die Kernkomponenten der Infrastruktur vor und wie Forschende die Infrastruktur für ihre Forschungsfragestellung einsetzen können.

[1] http://ec.europa.eu/research/infrastructures/pdf/esfri_brochure_0113.pdf

[2] http://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/roadmap_2006/esfri_roadmap_2006_en.pdf

[3] <http://www.cceh.uni-koeln.de/files/ThallerIntroWahn.pdf>

Schritte zur Integration einer Ressource in CLARIN-D

Thomas Eckart, Volker Boehlke, Jörg Knappen

In den Digital Humanities werden kontinuierlich neue Daten und Verfahren erstellt und aufbereitet welche in der jeweiligen Community einen hohen Stellenwert haben, aber auch für Fachanwender aus verschiedenen anderen Bereichen eine hohe Relevanz besitzen. Dabei treten erfahrungsgemäß wiederholt vergleichbare Probleme auf, welche häufig aus Zeitmangel gar nicht oder nur rudimentär adressiert werden. Hierzu zählen:

- die langfristige Archivierung und Wiederverwendbarkeit von Ressourcen und Werkzeugen
- die Bereitstellung von Metadaten und die Sichtbarkeit dieser in einem zentralen Metadatenkatalog
- die Bereitstellung von Methoden welche das sichere Zitieren einer Ressource und deren einzelner Bestandteile ermöglichen
- der feingranulare Zugriff auf Ressourcen mit dem Ziel der Wiederverwendbarkeit
- die Anbindung an ressourcen- und anbieterübergreifende Anwendungen, welche effizientes Arbeiten mit einer Vielzahl digitaler Ressourcen ermöglicht (Arbeit mit verschiedenen zugriffsgeschützten Ressourcen, Speicherung und Austausch umfangreicher Zwischenergebnisse, ...)
- die Bereitstellung von Dokumentationen sowie von Lehr- und Lernmaterialien zu digitalen Ressourcen und Werkzeugen

In diesem Vortrag, dem zweiten der Sektion, wird daher die Integration von Ressourcen in die CLARIN-D Infrastruktur thematisiert. Dabei werden die hierfür obligatorischen und optionalen Schritte zur Einbindung von Daten und Verfahren definiert und detailliert beschrieben. Die Basis einer jeden Integration von Ressourcen in CLARIN-D bildet dabei die Erzeugung geeigneter Metadaten im CMDI-Format und deren Bereitstellung über eine standardisierte Web-Schnittstelle. Auf dieser Basis bauen diverse Anwendungen (wie zum Beispiel der Metadaten-Katalog Virtual Language Observatory VLO) auf. Ein weiterer obligatorischer Bestandteil ist die Sicherstellung der Zitierbarkeit durch die Nutzung von globalen Identifikationsdiensten wie dem Handle System.

Weitere wichtige Bestandteile möglicher Integrationsmaßnahmen umfassen dabei:

- Jedes der neun CLARIN-D Zentren unterhält ein Repositorium für die langfristige Archivierung und Bereitstellung digitaler Ressourcen. Die Repositorien sind mit dem Data Seal of Approval zertifiziert.
- Bereitstellung Webservice-basierter Zugriffsmethoden auf Daten und Ressourcen. Hier wird die Grundlage für den granularen Zugriff auf Daten und Workflows und damit für die Möglichkeit der einfachen, gezielten und effizienten Wiederverwendung einer Ressource in neuen Kontexten gelegt.
- Nutzung der föderierten Authentifikations- und Autorisierungsinfrastruktur CLARIN-D AAI. Diese erlaubt es den Zugriff auf eine Ressource auf bestimmte Nutzer bzw. Nutzergruppen einzuschränken und verknüpft dies mit Methoden, welche Single-Sign-On Funktionalität ermöglichen.
- Die Anbindung an die CLARIN-D Federated Content Search stellt eine attraktive Möglichkeit der inhaltsbasierten Suche auf einer Vielzahl verschiedener textueller Ressourcen gleichzeitig dar.

- Das Problem der Speicherung von Zwischenergebnissen kann mit Hilfe der Anbindung an die CLARIN-D Workspaces gelöst werden.
- Die TeLeMaCo-Sammlung von Lehr- und Lernmaterialien erlaubt ein gezieltes Finden von Kurzanleitungen, Handbüchern sowie kleinen und großen Lerneinheiten.

Gesamtziel des Vortrags ist es den Teilnehmern zu verdeutlichen, welche neuen Möglichkeiten sich durch die oben genannten Formen der Integration bieten, mit welchen Methoden und Werkzeugen dabei gearbeitet werden kann und in welcher Form Dokumentation zu jedem dieser Schritte verfügbar ist. Zudem sollen bekannte Problemfelder, wie die Problematik der geeigneten Granularität der Metadaten und des Zugriffs auf eine Ressource diskutiert und exemplarische Lösungen vorgestellt werden. Die Teilnehmer sollen in die Lage versetzt werden, den für die Integration von Ressourcen notwendigen Aufwand korrekt abschätzen zu können. Es soll zudem verdeutlicht werden, welche Synergien und Potentiale durch eine solche Integration in eine Forschungsinfrastruktur wie CLARIN-D mit relativ geringem Aufwand möglich sind.

Webdienste und WebMAUS

Thomas Kisler

Die Common Language Ressources and Technology Infrastructure, CLARIN, ist eine Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften. Ziel der Infrastruktur ist es, sprachbasierte Ressourcen und Dienste dezentral anzubieten, sowohl für geschriebene als auch gesprochene Sprache.

Ein wichtiger Bestandteil von CLARIN sind Webservices, die es Wissenschaftlern ermöglichen, speziell auf ihre Anwendungsbedürfnisse hin entwickelte Softwarepakete im WWW zu nutzen, ohne diese auf dem eigenen Rechner installieren zu müssen. Ebenso können etablierte und weitverbreitete Softwaretools auf diese Weise um neue, webbasierte Dienste erweitert werden und somit neue Funktionalität anbieten. Ein besonderer Vorteil webbasierter Dienste besteht zudem darin, dass ohne Zutun der Nutzer die jeweils neueste Version der Software verwendet werden kann.

Der Zugriff auf Ressourcen gesprochener Sprache erfordert mindestens eine orthographische Transkription der Audiodateien; für weitere Analysen sind detailliertere Transkriptionen, etwa eine breite phonemische oder gar eine enge phonetische Segmentation und Etikettierung notwendig. Ein orthographisches Transkript kann in der Regel einfach und ohne besondere Kenntnisse erstellt werden, häufig ist es sogar, z.B. im Falle von Korpora gelesener Sprache, bereits verfügbar. Die Erstellung einer phonemischen Transkription oder einer phonetischen Segmentation dagegen erfordert Spezialkenntnisse und ist extrem zeitaufwendig – so kann eine enge phonetische Segmentation bis zu tausendmal so lang dauern wie die Äußerung lang ist.

Das Münchener automatische Segmentationssystem MAUS erstellt aus dem orthographischen Transkript einer Äußerung eine Segmentation auf Phonemebene, d.h. jeder Laut der Äußerung wird mit dem Sprachsignal aliniert. Die Besonderheit von MAUS ist, dass es aus dem orthographischen Transkript aufgrund statistischer Ausspracheregeln Aussprachehypothesen generiert und diese mit dem Sprachsignal abgeglichen werden; MAUS gibt als Ergebnis die wahrscheinlichste Aussprachehypothese zurück und kann damit die typischen Koartikulationsphänomene gesprochener Sprache berücksichtigen.

Im Kontext von CLARIN-D wurden eine grafische, webbasierte Benutzeroberfläche für MAUS sowie Programmierschnittstellen in Form von Webservices zum einfacheren Einbinden der Funktionalität in externe Anwendungsprogramme entwickelt. Diese Webservices sind, wie alle CLARIN-D Ressourcen, in CMDI-konformen Metadaten beschrieben und können somit von geeigneten Anwendungsprogrammen automatisch benutzt werden.

Die grafische Benutzeroberfläche im WWW erlaubt es einem Nutzer, interaktiv Audiodateien und dazugehörige orthographische Transkripte im Browser hoch- und nach Abschluss der Bearbeitung die Segmentationsdaten herunterzuladen.

WebMAUS unterstützt aktuell neun Sprachen und ist ihm Rahmen von CLARIN-D Showcases in bestehende linguistische Workflows integriert.

Motivation einer Sektion zu CLARIN-D und Zusammenfassung der Abstracts

Christoph Draxler
Bayerisches Archiv für Sprachsignale
Institut für Phonetik und Sprachverarbeitung
LMU München
draxler@phonetik.uni-muenchen.de

1. Übersicht

Für die 1. Jahrestagung der DHd planen wir eine Sektion zu CLARIN-D, der deutschen Forschungsinfrastruktur-Initiative für die Geistes- und Sozialwissenschaften mit dem Schwerpunkt auf Sprachressourcen. Diese Sektion besteht aus drei Vorträgen:

1. *CLARIN-D und Forschungsinfrastrukturen*: Thorsten Trippel (Universität Tübingen), Dieter van Uytvanck (Max-Planck Institut für Psycholinguistik, Nijmegen)
2. *Schritte zur Integration einer Ressource in CLARIN-D*: Thomas Eckart, Volker Boehlke (Uni Leipzig), Jörg Knappen (Universität des Saarlands)
3. *Web-Services und WebMAUS*: Thomas Kisler (LMU München)

2. Abstract

Die drei Vorträge bauen aufeinander auf: der erste ordnet CLARIN-D in die europäischen Infrastrukturinitiativen ein, der zweite beschreibt den Umfang der CLARIN-D Ressourcen und das prinzipielle Vorgehen der Integration von externen Ressourcen in die CLARIN-D Infrastruktur, der dritte zeigt exemplarisch einen für CLARIN-D entwickelten Webdienst.

Die europäische CLARIN Infrastruktur besteht aktuell aus acht Ländern sowie einer internationalen Organisation als Mitgliedern, und einem Land als Beobachter. Die deutsche CLARIN-D Initiative ist dezentral organisiert: neun Zentren decken die Erstellung und Pflege von sowie den Zugriff auf sprachbasierte Ressourcen und Dienste ab. In der aktuellen Implementierungsphase sind in CLARIN-D bereits wichtige Infrastrukturkomponenten erfolgreich implementiert worden, so z.B. das einheitliche Login, die Einrichtung von menschen- und maschinenzugänglichen Repositories an allen CLARIN-D Zentren, Unterstützung sowohl des automatischen Harvesten der Metadaten für den globalen Datenkatalog des Virtual Language Observatory (VLO) als auch eine erste Version der interaktiven verteilten Inhaltssuche in den Datenbeständen der Zentren. Darüberhinaus sind viele Textkorpora und Sprachdatenbanken CLARIN-konform aufbereitet und in die Repositories aufgenommen worden, und Tools für die web-basierte Bearbeitung der Daten entwickelt und implementiert worden.

Im zweiten Vortrag beschreiben wir, wie bislang verteilte und uneinheitlich aufgebaute Text- und Sprachressourcen so aufbereitet werden, dass sie CLARIN-D konform beschrieben und über ein einziges Login genutzt werden können. Dabei ist es zentrales CLARIN-D Anliegen, dass die Daten – inklusive aller Zugangs- und Nutzungsrechte – bei den bisherigen Eigentümern bleiben.

Natürlich besteht auch die Möglichkeit, Datenbestände an ein CLARIN-D Zentrum zu übertragen, z.B. um die dauerhafte Verfügbarkeit und Langzeitarchivierung zu sichern.

In diesem Vortrag werden die wichtigsten Schlüsseltechnologien wie die komponentenbasierte Metadaten-Architektur CMDI, die Verwendung von dauerhaft gültigen Persistent Identifiern, der zentrale Metadatenkatalog der VLO und die verteilte Inhaltssuche im Detail, die Verkettung linguistischer Verarbeitungstools im Web vorgestellt, und der Aufbau einer Dokumentation und Sammlung von Lehr- und Lernmaterialien vorgestellt.

Der dritte Vortrag beschreibt die Entwicklung von CLARIN-D Webdiensten und geht auf den für CLARIN-D implementierten Dienst WebMAUS ein. Konkret geht es um das in der Verarbeitung gesprochener Sprache notwendige, aber sehr zeitaufwendige Etikettieren und Segmentieren von Sprachdateien. Bei dieser Segmentation wird das Sprachsignal mit einer Wort- oder Lautfolge aliniert, so dass Analysen des und Suchen im Sprachsignal über den Inhalt der Äußerung möglich sind. Grundlage der Entwicklung von WebMAUS ist, dass eine einfache orthographische Transkription einer Äußerung relativ einfach und ohne spezielles phonologisches oder phonetisches Wissen möglich ist. Liegt eine solche orthographische Transkription vor, dann können in einem automatischen Verfahren daraus Aussprachehypothesen vorgeschlagen und diese dann mit dem Sprachsignal abgeglichen werden.

WebMAUS ist in die linguistische Verarbeitungskette von CLARIN-D eingebunden und hat sich in kurzer Zeit zu einem vielgenutzten Webdienst entwickelt. Zu den Anwendern zählen nicht nur Phonetiker und Linguisten, sondern zunehmend auch unter anderem die Ethnologie, Dialektologie, Sprachtechnologie und Kommunikationswissenschaften.

3. Vortragende

Thorsten Trippel: Seminar für Sprachwissenschaft, Universität Tübingen; Liaison zur europäischen CLARIN Forschungsinfrastruktur

Dieter van Uytvanck: Language Archives, Max-Planck Institut für Psycholinguistik, Nijmegen; Leiter der technischen Infrastruktur in CLARIN-D

Volker Boehlke: Abteilung Automatische Sprachverarbeitung, Uni Leipzig; Betreuung der CLARIN-D Facharbeitsgruppen

Thomas Eckart: Abteilung Automatische Sprachverarbeitung, Uni Leipzig; Repositories und Inhaltssuche

Jörg Knappen: Englische Sprach- und Übersetzungswissenschaft, Universität des Saarlands; Repositories und Inhaltssuche

Thomas Kisler: Bayerisches Archiv für Sprachsignale, LMU München; WebDienste

Unterstützung von Forschungsprozessen in einem internationalen Forschungsprojekt: Variantengrammatik des Standarddeutschen

Gunter Vasold, Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities (ZIM – ACDH), Universität Graz, Österreich

Einleitung

In diesem Poster wird anhand eines konkreten Forschungsprojekts eine verteilte Arbeitsumgebung vorgestellt, die es kollaborativ arbeitenden ForscherInnen ermöglicht, (Zwischen)Ergebnisse zu erfassen, iterativ anzureichern, zu verbessern und die geleistete Arbeit zu dokumentieren bzw. dokumentieren zu lassen. Bei der Entwicklung wurde besonderes Augenmerk auf die Unterstützung typisch geisteswissenschaftliche, d.h. hermeneutisch ausgerichteter Forschungsprozesse gelegt.

Das Projekt

Das Forschungsprojekt "Variantengrammatik des Standarddeutschen" ist ein durch den Schweizerischen Nationalfonds (SNF), die DFG und den österreichischen FWF gefördertes Projekt. Forschergruppen in Salzburg (ursprünglich Augsburg), Graz und Zürich untersuchen gemeinsam mit Kooperationspartnern in Liechtenstein, Belgien, Luxemburg und Südtirol systematisch nationale und regionale Unterschiede in der Grammatik der deutschen Standardsprache. Methodisch stützt sich das Projekt auf die Analyse eines Korpus mit Zeitungstexten aus allen deutschsprachigen Ländern, das mehr als 300 Millionen Tokens enthält.

Herausforderungen

Die geographische Verteilung der beteiligten ForscherInnen auf mehrere Länder erforderte Hilfsmittel zur Unterstützung kollaborativer Arbeitsweisen, die größtenteils mit bestehenden Werkzeugen abgedeckt werden konnten. So wurde der gemeinsame Zugriff auf das Korpus durch ein von Semtracks bereitgestelltes Webinterface zur Corpus Workbench (CWB) gelöst. Die projektinterne Kommunikation erfolgte in bewährter Weise über E-Mail und Skype; auch für die Literaturverwaltung standen zu Projektbeginn entsprechende netzwerkfähige Produkte zur Verfügung. Alleine für die Verzeichnung und Verwaltung bereits untersuchter grammatischer Varianten und die projektinterne Bereitstellung von Zwischenergebnissen fehlte es an fertigen Lösungen. Ein erster Ansatz, nämlich die benötigten Daten in einem via Google-Docs zentral bereitgestellten Tabellendokument zu speichern, erwies sich mit wachsender Datenmenge bald als unpraktikabel, nicht zuletzt, weil die zu speichernden Daten nur unzureichend in eine zweidimensionale Matrix abbildbar waren.

Die Lösung

Anfang 2012 wurde das Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities (ZIM – ACDH) mit der Lösung dieses Problems betraut. Es wurde auf Basis eines zuvor für ein anderes Projekt entstandenen Frameworks eine Datenbank mit Web-Interface entwickelt, über das alle ProjektmitarbeiterInnen die von ihnen untersuchten Varianten zentral erfassen und schrittweise weiter bearbeiten können. Dazu gehört die Dokumentation aller Korpusabfragen und daraus resultierender Ergebnisse ebenso wie das Speichern und Kommentieren statistischer Auswertungen, sowie das Zitieren von Belegstellen und von variantenspezifischer Sekundärliteratur. Jede Variante

kann grammatisch (Wortbildung, Flexion, Genus, Valenz/Rektion; teilweise mit mehreren Kategorieebenen) und inhaltlich (Einzelvariante, Variantenphänomen) kategorisiert und zusätzlich frei beschlagwortet werden.

Besonderes Augenmerk wurde dabei dem prozesshaften Voranschreiten der Arbeit geschenkt. Jeder einzelne Speichervorgang wird unter Berücksichtigung von BearbeiterIn, Zeitstempel, Bearbeitungsstatus und einer allfälligen Anmerkung protokolliert. Diese Information steht bei Bedarf im Kontext jeder einzelnen Variante zur Verfügung. Zusätzlich kann gezielt nach diesen Informationen gesucht werden. Die Prozessdaten bilden das Gerüst eines „Laborbuches“ und sind wesentlich für die Zuschreibung geleisteter Beiträge. Sie sind auch nützlich für das Projektmanagement, etwa zur Identifizierung von Varianten mit Bearbeitungsbedarf oder als Maßzahlen für die Projektplanung.

Ein formulargesteuertes Suchwerkzeug erlaubt kombinierte Abfragen über alle Datenbankfelder. Die Suche liefert eine tabellarische Überblickdarstellung, aus der heraus einzelne Varianten im Detail angesehen oder direkt zum Bearbeiten geöffnet oder gelöscht werden können. Als arbeitstechnisch hilfreich hat sich die Möglichkeit erwiesen, im Suchresultat bestimmte Varianten zur genaueren Analyse auswählen zu können. Die gewählten Einträge können in einer Vorschaufunktion rasch durchgesehen und aus dieser heraus wieder deseletiert oder nachbearbeitet werden. Aus erkenntnistheoretische Sicht ergibt sich daraus die Möglichkeit, aus im System vorhandenen Daten Erkenntnisse abzuleiten und diese in Form adaptierter oder neuer Daten unmittelbar wieder in das System zurückfließen zu lassen. Aus dem Suchergebnis heraus können gewählte Varianten auch in verschiedene Formate (Excel, PDF, HTML/XML) exportiert werden, was etwa ein Korrekturlesen auf Papier oder die Nutzung spezieller Analysesoftware ermöglicht.

Da die Datenbank nicht nur den Forschungsprozess unterstützen und dokumentieren soll, sondern vor allem die Datengrundlage für ein zum Projektende geplantes Handbuch darstellt, wurde zusätzlich eine Registersicht als Vorschau und Ausgangspunkt für die noch zu erstellende Druckvorlage implementiert.

Résumé

Das Beispiel der Forschungsdatenbank zur „Variantengrammatik des Standarddeutschen“ zeigt, dass mit geringem Aufwand individuelle VRE-artige Arbeitsumgebungen realisiert werden können, die auf pragmatische Weise die Verwaltung von Forschungsergebnissen im Sinne von (vorläufigen) Erkenntnissen erleichtern. Eine spezifisch geisteswissenschaftliche, d.h. hermeneutische Herangehensweise im Sinne des von Gadamer als hermeneutischer Zirkel beschrieben Erkenntnisprozesses erfordert, dass in einem solchen System nicht nur Ergebnisse, sondern auch deren Fortentwicklung und Veränderung abbildbar sind und entsprechende Forschungsprozesse unterstützt werden. Es geht dabei weniger um die Vorgabe genau spezifizierter Workflows, sondern um die flexible Unterstützung iterativer Prozesse und deren Dokumentation.

Literatur

<http://variantengrammatik.net/>

Christa Dürscheid, Stephan Elspaß und Arne Ziegler: Grammatische Variabilität im Gebrauchsstandard – das Projekt »Variantengrammatik des Standarddeutschen«. In: Marek Konopka et al. (Hgg.), Grammar und Corpora 2009, Tübingen 2011, S. 123–140.

Hans-Georg Gadamer: Vom Zirkel des Verstehens. In: Hans-Georg Gadamer, Gesammelte Werke 2, Tübingen 1993, S. 57–65 (Zuerst erschienen in: Festschrift für Martin Heidegger zum 70. Geburtstag, Pfullingen 1959).

Jan Potthoff und Sebastian Rieger: Elektronisches Laborbuch: Beweiswerterhaltung und Langzeitarchivierung in der Forschung. In: Silke Schomburg, Claus Leggewie, Henning Lobin und Cornelius Puschmann (Hgg.) Digitale Wissenschaft, Köln 2011, S. 149–156.

Thomas Süptitz, Stephan J. J. Weis und Torsten Eymann: Was müssen Virtual Research Environments leisten? – Ein Literaturreview zu den funktionalen und nichtfunktionalen Anforderungen. In: Proceedings of the 11th International Conference on Wirtschaftsinformatik (WI2013), Leipzig 2013, S. 327–341.

Das Balkanbild in Deutschland während der letzten 300 Jahre

- Eine digitale Plattform zu Analyse und Erschließung multilingualer Dokumente über die Balkanländer und das Osmanische Reich -

Cristina Vertan, Walther v. Hahn

Universität Hamburg

Nach offiziellen Statistiken¹ lebten 2011 in Deutschland mehr als 1 Million Menschen aus den Balkanländern und der Türkei. Diese Zahl wird sicherlich ab 2014 weiter steigen, wenn Rumänen und Bulgaren keinen Sonderregelungen auf dem Arbeitsmarkt mehr unterliegen werden.

Ein wichtiger Aspekt in der Debatte über die Integration dieser Menschen ist ihr kulturhistorischer Hintergrund: Alle Länder auf den Balkan waren fast 500 Jahre unter osmanischer Herrschaft.

Deutschland war zwar durch seine geographische Lage immer viel näher (kulturell und politisch) am Osmanischen Reich als die meisten westeuropäischen Länder; Deutsche Kaufleute in Siebenbürgen (also damals im Osmanischen Reich) sind hier nur ein Beispiel. Später im 18. und 19. Jh. sind die Kontakte intensiver geworden auch durch das Interesse von deutschen Künstlern und Wissenschaftlern am Erbe der Antike, das sich damals außer Italien durch die Herrschaftsverhältnisse natürlich überwiegend auf osmanischen Boden befand.

Erstaunlich ist aber, dass der Informationsfluss via originalen Dokumenten (aus der Feder westlicher Zeitgenossen des Osmanischen Reiches oder Reisender in dasselbe), eigentlich nur ein sehr dünnes Rinnsal war. Hauptsächlich sind es die Schriften des Universalgelehrten Dimitrie Cantemir. Er war der Sohn eines moldawischen Herzogs, war aufgewachsen als Gefangener in Istanbul und wurde später Mitglied der Preußischen Sozietät der Wissenschaften, der nachmaligen Akademie. In deren Auftrag dokumentiert er das erstes Mal das politische und soziale Leben im Osmanischen Reich in seinem Buch „Geschichte der Entstehung und des Verfalls des Osmanischen Reiches“ (Anfang des 19. Jhds.). Cantemir beschreibt viele historische Ereignisse aber erklärt auch die zentrale Verwaltung und das Militär, die Steuern in den von Osmanen eroberten Regionen, die Verwaltungsprinzipien für die unterschiedlichen Nationalitäten, usw. Das Werk hat er in Lateinisch verfasst, Jahre später wurde es übersetzt ins Deutsche, Englische, Französische, Rumänische und Russische. Diese Übersetzungen weichen aber vom Original ab und reflektieren in diesen Abweichungen die unterschiedlichen Kenntnisse und Sichten, die andere Länder über das Osmanische Reich hatten. Cantemirs Werk blieb bis Mitte der XIX Jh. die Hauptreferenz für alle Mitteleuropäer, die sich mit dem Thema „Osmanisches Reich“ befassten. Cantemir schreibt für die Preußische Akademie der Wissenschaften auch ein weiteres Werk „Die Beschreibung der Moldau“, das erste Werk über dieses Land am Rande des damaligen Europa. Das Buch enthält auch die erste detaillierte Kartographie des Herzogtums Moldawien.

¹ (<http://de.statista.com/statistik/daten/studie/1221/umfrage/anzahl-der-auslaender-in-deutschland-nach-herkunftsland/>)

Im Übrigen gibt es auch mehrere Reiseberichte durch die Walachei, Moldawien und das übrige osmanische Reich, die teilweise ein realistisches Bild über die Balkanhalbinsel in der Zeit vom 15. (nach der Eroberung Konstantinopels) bis zum 19. Jh. überliefern.

Die Frage „Was haben wir für ein Türkeibild erworben?“ scheint auf den ersten Blick gerade durch diese schmalen Traditionstränge eher leicht zu beantworten zu sein.

Aber alle diese Dokumente befinden sich in unterschiedlichen Bibliotheken, sie sind in unterschiedlichen Sprachen (Lateinisch, Englisch, Russisch, Französisch, Türkisch, u.a.) oder Sprachvarianten (osmanisches Türkisch, Frühneuhochdeutsch, Altrumänisch) verfasst, Sprachen, die für Laien, aber auch viele Wissenschaftler nur teilweise verständlich sind. Dazu kommen für jeden Nichtwissenschaftler Schwierigkeiten, diese Dokumente überhaupt zu verstehen, denn

- die Ortsnamen haben sich - zum Teil mehrfach - geändert,
- die Karten sind unrichtig oder entspringen politischem Wunschdenken,
- Personennamen sind anders transliteriert als heute und
- die Werke enthalten in der Übersetzung wieder weitere anderssprachige Zitate.
- zur Interpretation der Schriften bedarf es eines enorm breiten und ungewöhnlichen Fakten- und Begriffswissens.

Selbst Turkologen haben bis heute bisweilen Schwierigkeiten, in diesem 500jährigen multikulturellen und vielsprachigen Forschungsfeld mit wenigstens drei verschiedenen Schriften, verlässliche Interpretationen abzugeben. Journalisten und Politiker in der langsam beginnenden Beitrittsdebatte sind mit dem Lesen originaler Quellen aus dem 15. bis 19. Jahrhundert und damit dem Verstehen unseres Bildes von der Türkei und den Balkanländern überfordert.

Ziel dieses Projekts ist die Entwicklung eines digitalen Pool originaler Dokumente vom 17. – 19. Jh., ausgestattet mit computerbasierte Analyse-, Interpretations- und Suchmethoden für das Verstehen des gemeinsamen Kulturraums Südost-Europa.

Im Projekt werden Methoden von multilingualen Text-Mining benutzt für:

1 Wissensbasis-Extraktion (d.h. Bearbeitung von Cantemirs Schriften, in denen viele Wörter und Begriffe des Osmanischen Reiches definitionsartig erklärt werden.). Hier werden Technologien im Bereich „Vergleichbare bilinguale Korpora“ angewendet.

2 Verbindung der Texte mit unterschiedlichen Wissensquellen zum Zwecke des "Multilingualen Textmining":

- Wissen, das aus den Texten extrahiert wurde,
- modernes Wissen über die Türkei und die Balkanländer,
- Auflösung der Eigen- und Ortsnamen,
- geographische Information über die jeweilige Größe des türkischen Reichs und wechselnde Ländergrenzen auf dem Balkan,
- sprachliche und sprachgeschichtliche Erklärungen zu verschiedenen Sprachen,

- Übersetzungen einzelsprachlicher Passagen,
- historischen Informationen über die türkischen Herrscher und Statthalter ,
- sozialen und kulturellen Informationen.

Der Beitrag wird die Systemarchitektur sowie die gezielte Anwendung von multilingualen Textmining –Methoden in diesem Kontext erklären

Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert?

Georg Vogeler <georg.vogeler@uni-graz.at>

Historische Rechnungsdokumente scheinen auf den ersten Blick hervorragend geeignet für die computergestützte Reproduktion:

- Sie sind hoch strukturiert,
- sie enthalten große Mengen an Einzelinformationen, die individuell nicht immer von hohem Quellenwert sind, als Aggregation jedoch Bedeutung gewinnen,
- sie enthalten Zahlen, mit denen gerechnet werden kann.

In den Zeiten, in denen die Kontaktzone zwischen Informatik und Geisteswissenschaften stark von sozialhistorischen Interessen geprägt war, war die Verbindung entsprechend eng: Die Quantifizierung als Methode historischer Forschung war auch schon seit den 1960er Jahren von den Möglichkeiten geprägt, die Computertechnologien boten.¹ Aus den Rechnungen wurden individuelle Datenbanken und tabellarische Darstellungen erstellt. Als gemeinsamer „Standard“ hat sich dabei die Tabellenkalkulation (z.B. MS Excel) durchgesetzt. Dieser Ansatz übergeht wichtige Informationen des originalen Dokuments, die insbesondere bei mittelalterlichen und frühneuzeitlichen Rechnungen von Bedeutung sein können:

- Die jüngst erschienene kritische Edition der Stadtrechnungen von Luxemburg² zeigt, daß Buchführung auch eine wichtige sprachhistorische Quelle ist, die Auskunft über Orthographie, Vokabular und Fachsprachen geben kann. Eine digitale Ressource, die nur buchhalterische Informationen wie Beträge, Konten und stichwortartige Beschreibungen der Buchungen enthält, verringert ihren linguistischen Quellenwert signifikant.
- Rechnungslegung und ihre Verschriftlichung waren im Mittelalter ein Prozeß, der mehrere Schritte umfaßte: Vorbereitung der Rechnung aus informellen Dokumenten (Belege, vorläufige Rechnungsregister u.ä.), Zusammenstellung in einer Reinschrift, mündliche Rechnungslegung vor einer Rechnungskontrollinstanz oder zumindest vor einer Person, die mit Hilfe des Abacus die Rechenprozesse nachvollzog, Aktualisierung von Schuldposten im Zuge der Erstattung durch den Debitor, Umwandlung von Soll-

¹ Herlihy ##, ##Rom-Tagung 1970.

² Die Rechnungsbücher der Stadt Luxemburg, bearb. v. Claudine Moulin u. Michel Pauly, z.Zt. 6 Hefte, Luxemburg 2007 - 2012 (Schriftenreihe des Stadtarchivs Luxemburg ...).

Buchungen in Ist-Zahlungen insbesondere in der Steuerverwaltung und in Grundherrschaften. Mittelalterliche Rechnungen sind also weniger abgeschlossene Finanzdokumente als mehr oder weniger lebende Protokolle von finanztechnischen Operationen.

- Das Layout von mittelalterlichen und frühneuzeitlichen Rechnungen ist nicht stabil. Rechnungen der Verwaltung entwickelten sich im Reich z.B. von Protokollen in längeren Textblöcken zu komplexen tabularischen Formen.³ Frühe Formen der doppelten Buchhaltung beruhten auf der Position der Buchung auf der Seite.⁴ Die Entwicklung der visuellen Form ist ein Teil der Erforschung der Geschichte des Rechnungswesens.
- Rechnungsbücher sind eine gute Quelle für Alltagsleben, Sozialgeschichte und Sachkultur.⁵ Diese Quelleninhalte erfordern die Verbindung der Buchungen mit Taxonomien, die nicht nach buchhalterischen Kriterien aufgebaut sind: Berufsgruppen, Warengruppen, Bevölkerungsschichten etc. Die Erschließung der Rechnungstexte alleine nach finanztechnischen Gesichtspunkten muß dafür um andere Arten inhaltlicher Erschließung ergänzt werden.

Die eingeführte Lösung für das Problem unterschiedlicher Forschungsinteressen an den Rechnungsbüchern ist die Konzentration bei der Edition auf eine Perspektive. Während die erwähnten Editionen der Luxemburger Stadtrechnungen auf eine diplomatische Transkription Wert legten, edierte z.B. Richard Knipping die Kölner Stadtrechnungen in tabellarischer Form.⁶ Im Medienkontext des Buchdrucks war es wirtschaftlich nicht vertretbar, eine Rechnung sowohl als Volltexttranskription als auch tabellarisch zu drucken.

Auf dem aktuellen Stand der theoretischen Diskussion über die Digitale Edition scheint das Problem gelöst: Eine kritische Edition kann als digitale Resource mehrere Interpretationsschichten einschließen und dem Benutzer die Auswahl der für ihn nützlichen Präsentationsform überlassen. Eine überblicksartige Recherche nach digitalen Ressourcen von mittelalterlichen und frühneuzeitlichen Rechnungsbüchern zeigt jedoch, daß diese Editionen

³ Mark Mersiowsky: Die Anfänge territorialer Rechnungslegung im deutschen Nordwesten. Spätmittelalterliche Rechnungen, Verwaltungspraxis, Hof und Territorium (zugl. Diss. phil. Münster 1992), Sigmaringen 2000 (Residenzenforschung 9); Georg Vogeler: Spätmittelalterliche Steuerbücher deutscher Territorien, Teil 1: Überlieferung und hilfswissenschaftliche Analyse, in: AfD 49 (2003), S. 165-295, Teil 2: Funktionale Analyse und Typologie, in: AfD 50 (2004), S. 57-204.

⁴ Federigo Melis: Storia della Ragioneria. Contributo alla conoscenza e interpretazione delle fonti più significative della storia economica, Bologna 1950; Franz-Josef Arlinghaus: Bookkeeping, Double-Entry Bookkeeping, in: Medieval Italy. An Encyclopedia, hg. v. Christopher Kleinhenz, New York 2004, S. 147-150; Basil S. Yamey: Scientific Bookkeeping and the Rise of Capitalism, in: EHR N.S. 1 (1949), S. 99-113.

⁵ Z.B. Gerhard Jaritz, Die Reimer Rechnungsbücher (1399-1477) als Quelle zur klösterlichen Sachkultur des Spätmittelalters, in: Die Funktion der schriftlichen Quellen i.d. Sachkulturforschung (Veröffentlichungen des Inst. f. mittelalterliche Realienkunde Österreichs 1), Wien 1976, S. 145-249; ##

⁶ Richard Knipping: Die Kölner Stadtrechnungen des Mittelalters mit einer Darstellung der Finanzverwaltung, 2 Bde., Bonn 1897 - 1898 (Publikationen der Gesellschaft für rheinische Geschichtskunde 15).

das theoretisch formulierte Potential jeweils nur eingeschränkt realisieren: Sie bieten Bilder zu den Texten, sie erlauben eine auf Taxonomien aufgebaute Suche oder sie transferieren die Informationen in eine Datenbank.⁷ Nur die Edition der Rechnungsbücher des *Royal Irish college of Saint George the Martyr* in Alcalá⁸ gibt dem Benutzer nicht nur Zugriff auf Bilder von Transkriptionen der Rechnung, sondern ermöglicht es ihm auch, mit den Buchungsposten zu rechnen.

Der vorgeschlagene Vortrag versucht einige Gründe für diese Situation zu analysieren, die sich aus grundlegenden Interessenslagen von verschiedenen Forschergruppen ergeben. Die Analyse zeigt damit auch, wie domänen spezifische Traditionen die produktive Zusammenarbeit zwischen Geisteswissenschaften und Informationstechnologen behindern, ja sogar die Digitalen Geisteswissenschaften als Vermittler zwischen beiden Gruppen aus ihren eigenen Forschungstraditionen heraus den Brückenschlag behindern. Der Vortrag sieht zentrale Gründe für die fehlende Umsetzung des Potentials digitaler Editionen für Rechnungsschriftgut in der Dominanz philologischer Editionsmethoden auch in der Forschungsdiskussion über die digitale Edition. So ist die TEI als de-facto-Standard digitalen Editierens z.B. an mehr an komplexen Überlieferungsverhältnissen, kodikologischen und paläographischen Details oder an linguistischen Phänomenen interessiert als an der Erschließung von Inhalten. Syd Baumann und Kathryn Tomasek haben deshalb erste Vorschläge erarbeitet, wie die TEI zu erweitern wäre, um Finanztransaktionen beschreiben zu können.⁹ Als alternativer Standard für die Kodierung von buchhalterischen Informationen steht mit XBRL (eXtended Business Reporting Language)¹⁰ ein flexibler Vorschlag von wirtschaftswissenschaftlicher Seite zur Verfügung, der mit der ‚Global Ledger‘-Taxonomie auch Kategorien für die Kodierung historische Buchhaltung bereit stellt. Es stellt sich die Frage, wie beide Standards mit einander in Verbindung gebracht werden können.

⁷ Regensburg Cameralia; Les comptes des consuls de Montferrand (1273–1319) 2006 (Éditions en ligne de l’École des Chartes, volume 16), <http://elec.enc.sorbonne.fr/montferrand/>, éd. R. Anthony Lodge 2006; Comptes de châtellenies, <http://www.castellanie.net>; Comédie-Française Register Project,

<http://web.mit.edu/hyperstudio/cfr/>; Open domesday. The first free online copy of Domesday Book, ed. by Anna Powell-Smith, J.J.N. Palmer, Univ. of Hull <http://www.domesdaymap.co.uk/>; Henry III fine rolls Project, King’s College London et al., 2007-2011 <<http://www.finerollshenry3.org.uk/>>, Die mittelalterlichen Schuld- und Rechnungsbücher des Deutschen Ordens um 1400. Eine synoptische Edition im Internet, ed. by Christina Link u. Jürgen Sarnowsky, Hamburg 2008 www.schuredo.uni-hamburg.de.

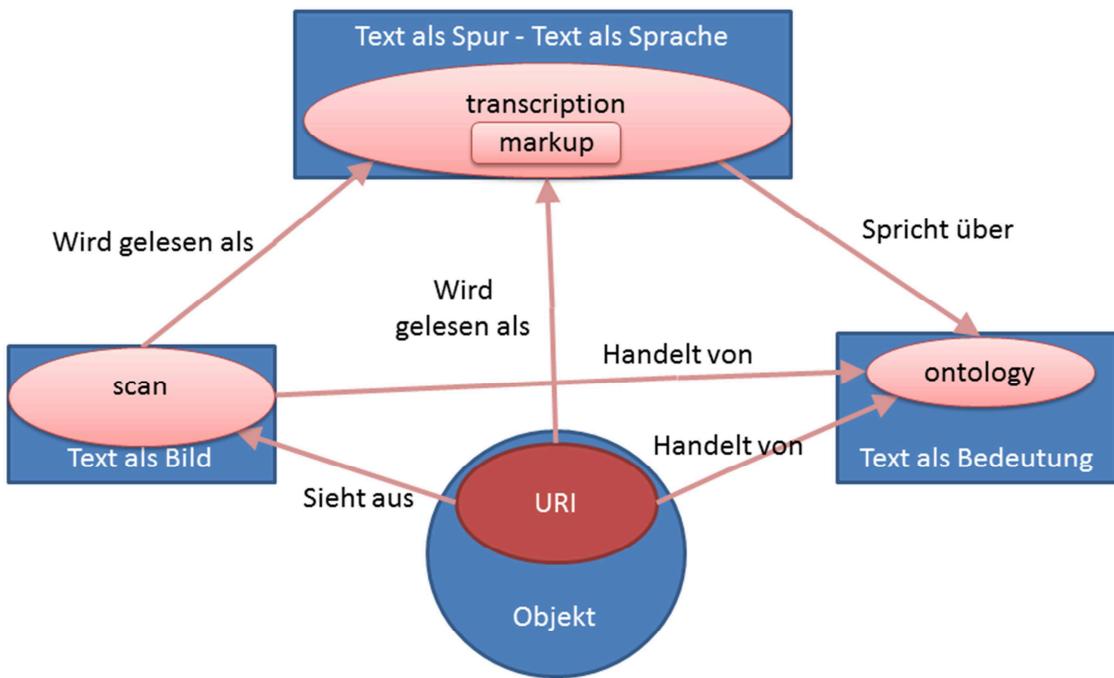
⁸ The Alcalá account book project, National University of Ireland, Maynooth 2008 <http://archives.forasfeasa.ie/>.

⁹ Syd Bauman, “Transactionography Customized Documentation,” *Encoding Historical Financial Records Open Access Library*, accessed October 28, 2013, <http://omeka.encodinghfrs.org/items/show/5>; Encoding Financial Records for Historical Research, paper presented at the TEI-MM 2012 in College Station, <http://idmc.tamu.edu/teiconference/program/papers/#encfin>; Encoding Financial Records for Historical Research, Whitepaper 2012, NEH Ref: HD-51224-11 <http://omeka.encodinghfrs.org/items/show/4>.

¹⁰ eXtensible Business Reporting Language, <<http://www.xbrl.org/>>, Global Ledger Taxonomy: <http://www.xbrl.org/GLTaxonomy>

Ausgehend von den Überlegungen, die Manfred Thaller 2012 präsentierte,¹¹ erscheint es angemessen, die verschiedenen Bedeutungsebenen der Rechnungsdokumente in einem RDF-Modell abzubilden, das den Text als Bild, als Spur als Sprache und als Bedeutung repräsentiert, und damit für unterschiedliche Forschungsfragen zugänglich macht.

Integration von Inhalt in ein Modell digitaler kritischer Editionen



Ein solches Modell kann mit Hilfe der *feature structures* der TEI serialisiert werden und mit einfaches XSLT in explizites RDF verwandelt werden, das sowohl auf Strukturen des XML-Dokumentobjektmodells als auch auf den in der TEI vorhandenen Möglichkeiten zur Repräsentation von Taxonomien aufbaut.

Das Kodierungsmodell ist theoretisch konsequent. Es wird gegenwärtig in verschiedenen Projekten auf seine praktische Umsetzbarkeit getestet und dabei nach Methoden gesucht, mit denen der Computer die Erzeugung einer digitalen Repräsentation des Dokuments „Rechnung“ z.B. mit Hilfe von automatischen Umrechnungen, automatischer Strukturerkennung und Pflege von Taxonomien und kontrollierten Vokabularen. Der Vortrag wird erste Ergebnisse als Proof of Concept präsentieren¹² und Lösungsvorschläge für die pragmatische Umsetzung in den Arbeitsalltag machen. Er soll damit dazu beitragen, die

¹¹ Manfred Thaller: What is a text within the Digital Humanities, or some of them, at least? *digital humanities 2012 (Hamburg)* <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/beyond-embedded-markup/>.

¹² Vgl. <http://gams.uni-graz.at/rem>

theoretischen Überlegungen aus dem Bereich der digitalen kritischen Edition, wie die verschiedenen Forschungsinteressen an Rechnungsschriftgut verbunden werden können, in ein realisierbares Modell digitaler Edition von mittelalterlichen und frühneuzeitlichen Rechnungsbüchern zu übersetzen.

„Von der Kurve zur Epoche“

Wie Rotquantitäten in Bildern Aussagen über Epoche, Genre und Stil erlauben
Ansätze einer computergestützten Bildanalyse

Neu: Das *Redcolor-Tool* aus dem Ommer-Lab der Ruprecht-Karls-Universität Heidelberg, konzipiert in Kooperation mit dem Kunsthistorischen Institut der Ludwig-Maximilians-Universität München

Proposal zum Vortrag auf der Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd)
eingereicht von Waltraud v. Pippich, Institut für Kunstgeschichte, Ludwig-Maximilians-Universität München, eingereicht am 1. 1. 2014

Ein neues Informatik-*Tool* für die Geisteswissenschaft

Das im Lab der *Computer Vision Group* am Lehrstuhl für Informatik an der Ruprecht-Karls-Universität Heidelberg unter der Leitung von Björn Ommer in Kooperation mit dem Kunsthistorischen Institut der Ludwig-Maximilians-Universität München eigens für das Forschungsvorhaben entwickelte Computerprogramm *Redcolor-Tool* soll im Vortrag erstmals der Öffentlichkeit präsentiert werden. Das Programm ermöglicht erstmals genauere Untersuchungen und umfassendere Forschungen zu Farbverläufen in Bildern. Dies wird ermöglicht durch digitale Datenanalyse. Im Zentrum steht die Frage: wie viel Rot ist im Bild? Für die am Kunsthistorischen Institut der Ludwig-Maximilians-Universität München unter der Betreuung von Hubertus Kohle entstehende Dissertation werden Rotquantitäten in Bildern gemessen. Das Computerprogramm *Redcolor-Tool*, ein Analyseinstrument der Informatik, liefert die Datenergebnisse. Die Daten beschreiben den Rotverlauf in Bildern sowie den Prozentsatz der anhand einer Skala auszuwählenden Schattierung von Rot bemessen auf die Gesamtfläche des analysierten Bildes. Die Daten werden aufbereitet in einem Graph als mathematische Funktion. In weiteren Schritten ermöglichen die durch das neue *Tool* verfügbaren digitalen Methoden die Vergleiche großer Corpora an Bildern.

Projektbeschreibung

Untersucht wurden eine sehr große Anzahl von Bildern: beginnend bei den Herrscherportraits des Absolutismus, hin über die (besonders zur Jahrhundertmitte) prekären Herrscherbildnisse des 19. Jahrhunderts, über zeitgenössische Darstellungen von Politikern auf dem roten Teppich, die z. Bsp. der Tagespresse entnommen wurden. Untersucht wurden auch Abbildungen

klerikaler Zeremonien wie Papstwahl oder Papstrücktritt, die Darstellungen entstammen den Feuilletons der internationalen Presse.

Was kann das *Redcolor-Tool* für eine an der Ästhetik der Bilder orientierte Wissenschaft leisten?

Das Computerprogramm *Redcolor-Tool* ermöglicht erstmals eine genauere Erfassung von Farbverläufen in Bildern via digitaler Datenanalyse. Dabei berechnet das *Tool* die Proportion der Rotpixel im Verhältnis zur Gesamtfläche des Bildes.

Orientierte man sich bei einer Unterscheidung der „Logik“ des Bildes von der Logik der Sprache an dem Jenenser Philosophen Gottfried Gabriel („Die Logik des Bildes bemisst das Verhältnis eines Teiles zum Ganzen. Die Logik der Sprache erfasst Unter- und Überordnungsverhältnisse.“, siehe z. Bsp. die Darstellungen in: Gottfried Gabriel, Logische Präzision und ästhetische Prägnanz, in: Literaturwissenschaftliches Jahrbuch N.F. Hg. von Volker Knapp, Kurt Müller, Klaus Ridder et al. Bd. 51, Berlin 2010, S. 375-390), ließe sich das neu gewonnene Verfahren als genuin bildlich arbeitende Methode begreifen. Wie groß ist der Teil roter Pixel im Verhältnis zum Gesamt des Bildes? Ist es viel? Ist es wenig? Was heißt „viel“ in dieser Epoche, was heißt „viel“ in diesem Genre (z. Bsp. Porträtmalerei, Schlachtenmalerei, Politikerdarstellung)? Und: welches Rot wird den vergleichenden Studien zu Grunde gelegt? Das Programm *Redcolor-Tool* stellt anhand einer umfassenden Rotskala besonders viele Rotwerte zu Analysezwecken zur Verfügung, wie im Vortrag anhand von Fallstudien gezeigt werden soll. Innerhalb einer an der Ästhetik der Bilder orientierten Wissenschaft liefert das *Recolor-Tool* einen Beitrag, den Eigenwert des Ästhetischen jenseits ikonographischer Fragestellungen zu erfassen und z. Bsp. die in kunsttheoretischen Fragestellungen kaum jemals ausführlich behandelten Eigentümlichkeiten von Farbe im Allgemeinen und der Farbe Rot im Besonderen zu behandeln. Es entsteht dabei eine neue Ästhetik, die Ästhetik der Linien der Kurvenfunktionen (siehe die Abbildung unten). Ohne den Zwischenschritt über die Sprache bildet das *Redcolor-Tool* die Proportion der Farbe ab und bietet ein Wahrnehmen pikturaler Größenverhältnisse durch die Zuordnung der Werte der x-Achse auf die Werte der y-Achse. Bei der Analyse großer Bildcorpora und dem Vergleich der Rotwerte verschiedener Bilder öffnet sich die Formanalyse hin zu historischen, kulturellen Fragestellungen, bildet gleichsam deren Basis. Im Ablesen der Graphen vollzieht sich eine neue Ästhetik, die eigentlich mit dem Gehalt der Bilder koinzidieren kann, wie zu zeigen sein wird.

Was leistet das *Redcolor-Tool* für die traditionelle kunstgeschichtliche Forschung als einer historisch orientierten Wissenschaft?

Im Nachfahren der Linien der Graphen und im Auffassen und Bewerten der Graphendaten vollzieht sich die Arbeit mit dem *Redcolor-Tool*. Die historische Perspektive arbeitet entlang der Frage, ob zu allen Zeiten alle Graphenverläufe denkbar sind, welchen Wandel die Rotwerte im Laufe der kunstgeschichtlichen Entwicklung nehmen. Die These ist, dass nicht zu allen Zeiten alle Farbverläufe denkbar sind und sich dieses implizite Wissen in den vom *Tool* bereitgestellten Daten wiederspiegelt. Von der informatisch berechneten, mathematischen Kurve hin zur historischen Epoche, „Von der Kurve zur Epoche“, gelangt die Arbeit des historisch orientierten Geisteswissenschaftlers mit dem neuen Informatikwerkzeug. Was wäre, wenn die Kurven Aussagen über die Entstehungszeit der Kunstwerke zuließen? Im Rahmen des *big data* Paradigmas, das für die Kunsthistorische Forschung zur Zeit z. Bsp. von Lev Manovich fruchtbar gemacht wird, lassen die vom *Redcolor-Tool* bereit gestellten Daten weitere Einblicke und Forschungen kulturgeschichtlicher Art erwarten.

Die eigentümliche Verbindung von Instrumenten der Informatik mit ästhetischen, hier quantitativen Fragestellungen und kulturellen Fragestellungen lässt Hoffnungen zu, den Zugang zu *implicit patterns* zu finden, die für die geisteswissenschaftliche Forschung erst durch große Datenmengen sichtbar werden könnten und der Auswertung und Einordnung harren.

Einer historisch argumentierenden, stilgeschichtliche Prozesse berücksichtigenden Wissenschaft wie der traditionellen Kunstgeschichte dient das *Redcolor-Tool* als Instrument zur Gewinnung zusätzlicher Informationen über die Kunstwerke. Die gewonnenen Daten können bestehendes kunsthistorisches Wissen ergänzen und von völlig neuen Seiten beleuchten.

Warum ergibt der elegante Absolutismus die eleganteste Linie?

Bei der Arbeit mit dem von der Heidelberger *Computer Vision Group* konzipierten Instrument stellte sich zum Beispiel die verblüffende Erkenntnis heraus, dass die Glanzzeit der Herrscherapotheose, der Absolutismus mit seiner hohen Eleganz, Pracht und wirkmächtiger Ikonographie die eleganteste Linie im Graphen produziert. Dies soll im Vortrag anhand des Bildnisses König Ludwig XIV. von Hyacinthe Rigaud erläutert werden.

Welche Potentiale stecken weiterhin in dem *Tool*, was ist neu?

Das Programm bietet auch die Möglichkeit einer 3-dimensionalen Darstellung der Graphen. Zu fragen wäre, ob die 3-dimensionale Aufbereitung der Proportionswerte der Eigenheit der Bilder gerechter wird als die 2-dimensionale Darstellung. Reagiert das Programm auf die häu-

fige Übermalung einer Stelle durch den Maler? An dieser Stelle ist die Leinwand plastischer, reicht an das 3-dimensionale heran. Das Analysetool stellt jedenfalls neue Fragestellungen in Aussicht, die besonders mit dem Vortrag von Katja Kwastek „Vom Bild zum Bild. Digital Humanities jenseits des Texts“ zusammenklingen.

Welche bislang während der bildwissenschaftlichen Forschung mit dem *Tool* aufgetretenen Probleme oder Schwierigkeiten werfen ein Licht auf die Perspektivität der unterschiedlichen Disziplinen der Informatik und der Bildwissenschaft?

Verschiedene disziplinspezifische Fragestellungen lassen sich entlang der Konzeption vom Programm *Redcolor-Tool* explizieren. Weshalb entwirft das Team der Heidelberger *Computer Vision Group* die x-Achse der Funktion als absteigenden, nicht als ansteigenden Rotwert? Gehen dabei Datensätze verloren? Unter welchen Umständen ließe sich von einem Programmierfehler sprechen? Handelt es sich um eine in der Sprache der Mathematik „ein-eindeutige“, umkehrbare Funktion, bei der jedem x-Wert ein y-Wert zugeordnet wird? Welche Konsequenzen haben diese Fragen für eine am Bild orientierte Forschung?

Eröffnen sich durch die Perspektiven der Bildwissenschaft auch Chancen für den Fortschritt im Fachbereich der Informatik? Wie ließen sich die traditionellen Ansätze erweiternde Forschungsagenden formulieren? Aus einer Verbindung Informatik - Kunstgeschichte - formal-ästhetische Bildwissenschaft entspringen, zählte man detailliert Schwierigkeiten bei der Konzeption des *Redcolor-Tools* auf (numerische Aufteilung etc.), Fragestellungen und Anforderungen an die Informatik, die sich produktiv auf diese Disziplin zurückwenden. Hinzu kommt: Was ist eigentlich das Bildhafte, Abbildende an den mathematischen Funktionen?

Wie geht es weiter? Was sind die nächsten denkbaren Schritte?

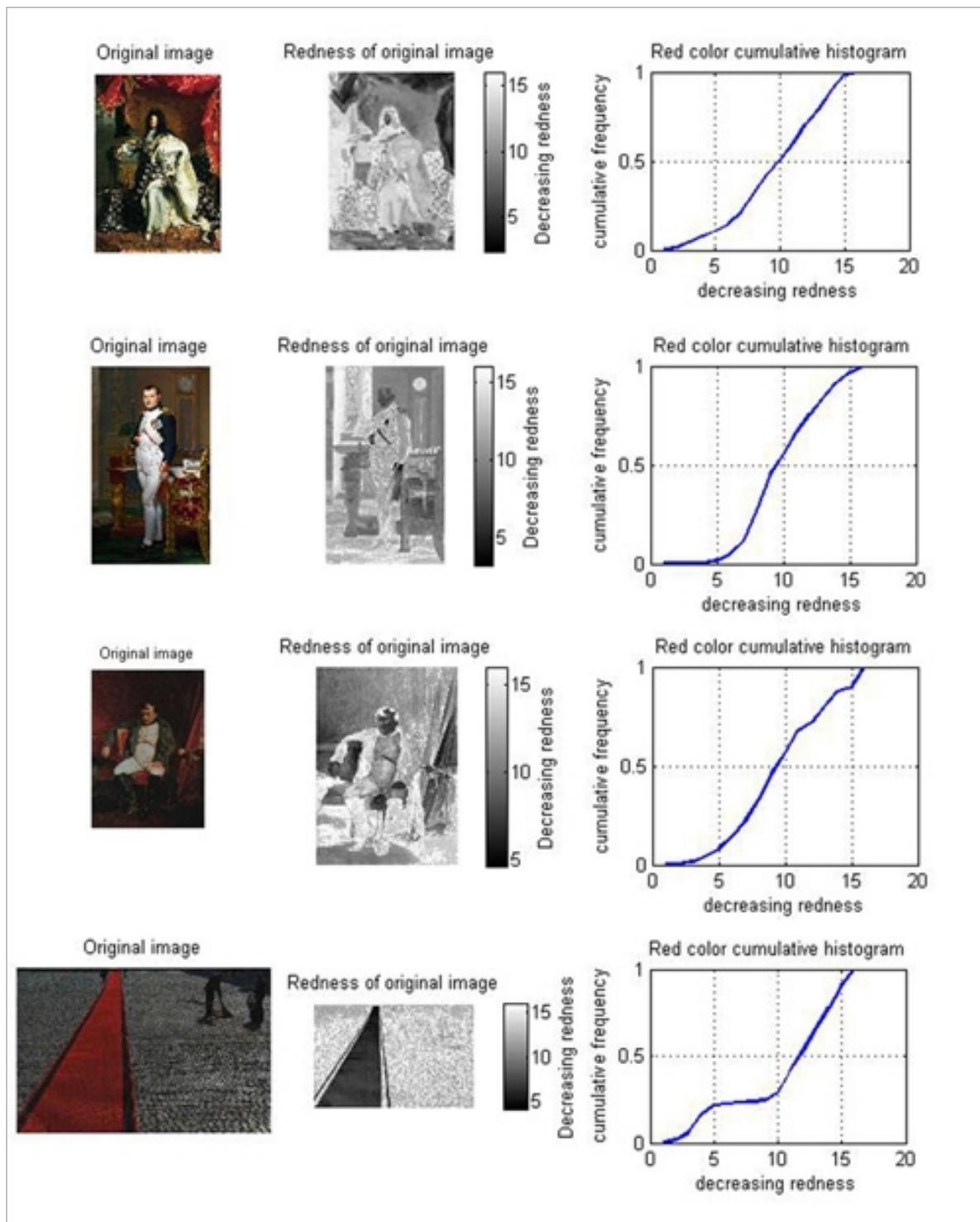
Das *Tool* sollte auch für Schwarzwerte, Weißwerte, für die Farbe Blau und viele weitere Farben konzipiert werden. Programme sollten geschrieben werden, die die Differenzwerte der Farb-Proportionen einzelner Bilder, möglicherweise auch von Bildserien, visualisieren.

Zum Schluss: Thesen zur Farbe Rot

Freilich soll neben den forschungstheoretischen, methodischen Fragestellungen im Vortrag auch die Theoriegeschichte der Farbe Rot berücksichtigt werden und Thesen zur Spezifik von Rot vorgestellt werden. Rot verkörpert Präsenz, Aktion, Bewegung, Leben, Blau hingegen Transzendenz, Ferne, Sehnsucht.

Gerne würde ich dem Publikum meine Forschungsergebnisse präsentieren. Ästhetische, kulturelle, historische Aspekte werden durch die Arbeit mit dem *Redcolor-Tool* berührt.

Abbildung





Verbundprojekt ‘MayaArch3D’

Ein webbasiertes 3D-GIS zur Analyse der Archäologie von Copan, Honduras

Jennifer von Schwerin, DAI/KAAK

Bisher gab es keine Infrastrukturen für die Aufbewahrung und Nutzung von 3D-Modellen, die in zunehmender Zahl von archäologischen Funden (z. B. Keramik, Skulpturen, Gebäuden oder ganzen Städten) angefertigt werden. Wo sollen solche digitalen Objekte aufbewahrt werden, damit Forscher sie im Internet sehen, analysieren und mit anderen Modellen vergleichen können? Wie können sie mit anderen archäologischen Daten verknüpft werden, damit man diese 3D-Modelle in einer virtuellen Welt erforschen kann?

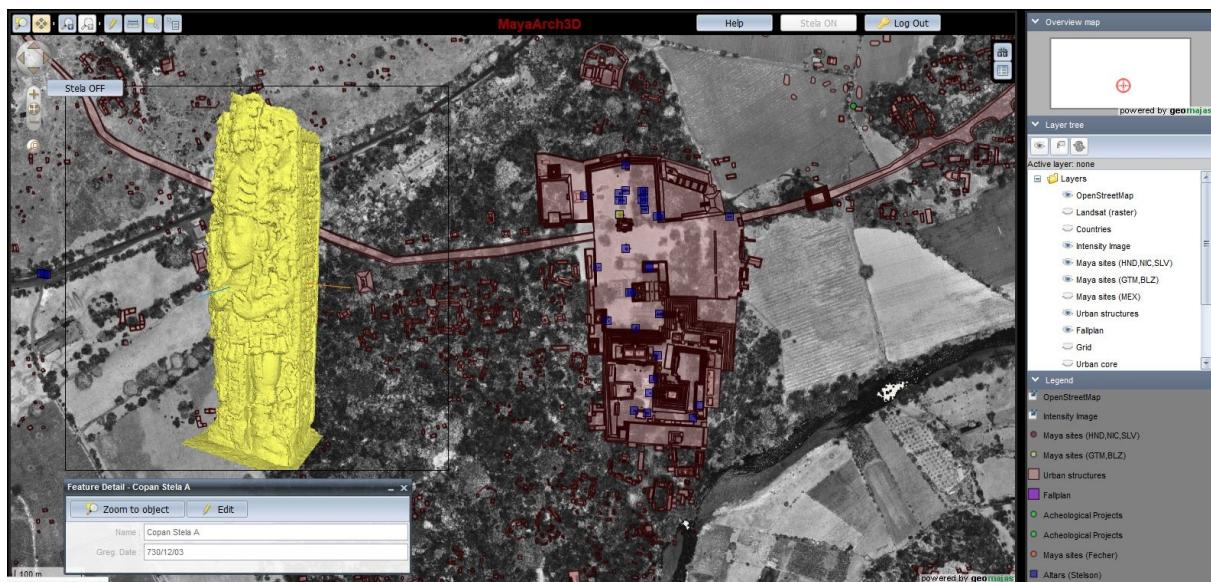


Abb. 1: 3D Geobrowser des MayaArch3D Projektes (<http://MayaArch3D.org>)

An dieser Schnittstelle zwischen Archäologie und Computerwissenschaften arbeitet das Verbundprojekt „MayaArch3D“, an dem das Deutsche Archäologische Institut (DAI), welches die Aufgaben der Altertumswissenschaften bearbeitet, und der Lehrstuhl für Geoinformatik und das Interdisziplinäre Zentrum für Wissenschaftliches Rechnen der Universität Heidelberg, welches sich mit den Fragen der Informatik beschäftigt, beteiligt sind. Ziel des Verbundprojektes MayaArch3D ist es, ein neues Forschungswerkzeug für die Archäologie zu entwickeln, welches es erlaubt, 3D-Modelle und Funktionen von Geographischen Informationssystemen (GIS) für die Dokumentation und Analyse archäologischer Fundstätten auf einer Internet-Plattform zusammenzuführen. Zweidimensionale und dreidimensionale Daten und Modelle von unterschiedlichster Art und Auflösung werden in einer sogenannten Geodateninfrastruktur (GDI) mit webbasierten interaktiven Analyse- und Visualisierungswerkzeugen eingebunden, so dass archäologische Analysen online in einem georeferenzierten System vorgenommen werden können. Abfragen von Zugangsmustern und Sichtverbindungen, Siedlungsplänen, topographischen Merkmalen, Richtungen, Material- und Fundverteilungen, die bisher nur in 2D-, oder 2.5D-Ansichten durchgeführt werden, sollen in der 3D-Umgebung möglich sein. Dabei werden weit verstreute Informationen und Objekte auf einer Internetplattform nach internationalen Standards dokumentiert, georeferenziert, virtuell zusammengeführt und analysiert.

Als Untersuchungsgebiet wurde die UNESCO Weltkulturerbestätte Copan in Honduras gewählt. Bei der in MayaArch3D entwickelten Infrastruktur handelt es sich jedoch um einen Prototyp, der nach

entsprechender Anpassung auch in anderen komplexen Ruinenstätten weltweit eingesetzt werden kann und eine Kombination aus Visualisierungs- und Analysewerkzeuge für die eHumanities bereitstellt.

Copan ist einer der wichtigsten Fundorte der klassischen Maya-Kultur. Er liegt nahe der Grenze der heutigen Staaten Guatemala und Honduras, innerhalb der südöstlichen Peripherie des Mayagebietes. Copan zeichnet sich gegenüber anderen Mayafundorten durch seine zahlreichen Tempel mit skulptierten Steinmonumenten und seiner besonders großen Anzahl an Hieroglypheninschriften aus. Seit 1885 werden in Copan Ausgrabungen durchgeführt. Archäologen aus Honduras, den USA, aus England, Japan, Frankreich und Deutschland konnten die ununterbrochene Geschichte eines Königreiches rekonstruieren, das zwischen 427 und 820 n. Chr. von sechzehn Herrschern regiert wurde.

Die Analyse der räumlichen Struktur von Copan kann wichtige Aufschlüsse über die Strukturierung des Ortes, die sozioökonomischen Verhältnisse, die Veränderung des Stadtbildes im Laufe der Zeit und die Geschichte der Maya-Kultur im Allgemeinen geben. Dafür werden Informatik-gestützte Werkzeuge benötigt, die im Rahmen des Projekts erstmals sowohl web-gestützt, als auch in 3D und auf Standards basierend erarbeitet werden sollen. Durch die Nutzung virtueller Landschaften und GIS-Karten, die mit durchsuchbaren Datenbanken verlinkt sind, können Wissenschaftler interaktive Analysen von Beziehungen und Veränderungen über Raum und Zeit anstellen.

Projektaktivitäten 2012-2013

Teilprojekt Archäologie

Die zentrale Aufgabe des Teilprojektes Archäologie ist es, die Informationen zur Archäologie, Architektur und zu Funden aus Copan nach Datentyp, Informationsgehalt und insbesondere hinsichtlich ihrer späteren Verwendung für die GIS-Analyse zu sammeln, zu strukturieren und in einer Datenbank abrufbar zu machen. Das Teilprojekt Archäologie koordiniert die Datensammlung von Informationen zum Copan des 8. und 9. Jahrhunderts in Archiven, Museen und im Feld, sowie die Datenaufbereitung und -strukturierung.

Am Anfang des Projektes im Jahre 2012 entwickelte das Teilprojekt Archäologie das konzeptuelle Design der Datenbank und bestimmte die Funktionen, die für das Tool (QueryArch3D) entworfen werden müssen, um ab 2014 die ersten Tests durchzuführen. iDAI.field – das Datenbanksystem des DAII wurde von den MayaArch3D Projektmitarbeitern in Zusammenarbeit mit der IT-Abteilung des Deutschen Archäologischen Instituts (DAI) Berlin unter der Leitung von R. Försch für die Verwaltung von Metadaten von 3D Modellen und den Einsatz in der Maya-Archäologie angepasst.

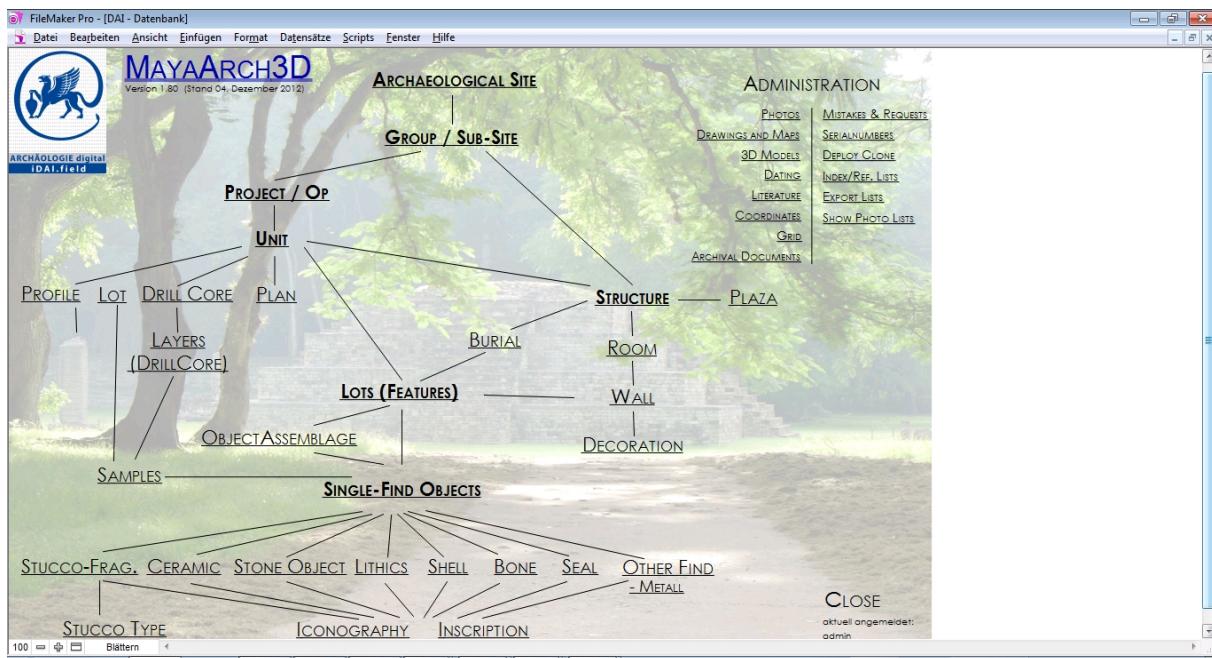


Abb. 3: iDAI.field adaptiert für Maya-Archäologie

Abb. 4: Neue iDAI.field Tabelle für 3D Modelle die von MayaArch3D angelegt wurde

Neue Tabellen für 3D Modelle, Fundstellen, Projekte, Sub-Projekte, Ikonographie, und Inschriften wurden angelegt. Die Benutzeroberfläche wurde neu gestaltet und auf Englisch und Spanisch übersetzt. Diese Datenbank dient als Zwischenlösung für die Datensammlung und Dateneingabe, bis 2014 ein leistungsfähigeres PostGris Datenbanksystem entwickelt wird.

Während Laborarbeiten in Bonn, einer Feldkampagne im April 2013 und einem Forschungsbesuch im American Museum of Natural History in New York im Mai 2013 wurden Daten für das System gesammelt, digitalisiert, und für die Eingabe in die Datenbank strukturiert.



Abb. 5: Altar der Stele 13 in Copan (KAAK, Laura Stelson)

Abb. 6: Keramik aus einer Altgrabungen im Gebäude 10L-18 (KAAK, Mike Lyons)

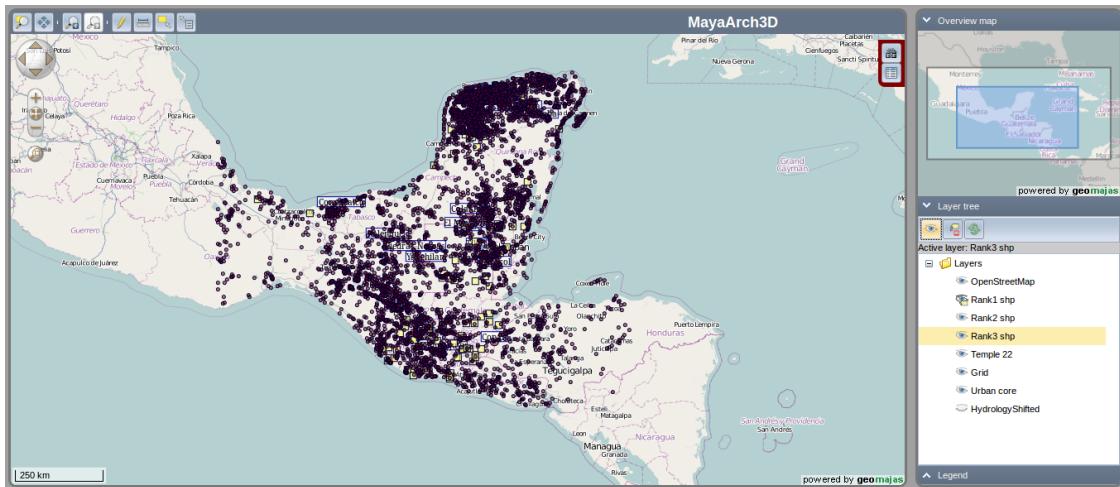


Abb. 7: Geobrowser mit 6000 archäologische Fundstellen

Die bisher gesammelten Daten umfassen Koordinaten von 6000 archäologischen Fundstellen, 224 davon in Honduras. Für Copan wurden ein geografisches Informationssystem (GIS) und eine 3D-computersimulierte Stadtlandschaft, die 24 Quadratkilometer und 3000 Strukturen des Copantals erfasst, erarbeitet. Derzeit werden weitere Daten in die Datenbank eingetragen: Struktur- und Gruppennamen, Typenbezeichnungen, Gebäudehöhen, Konstruktionsdaten und dazugehörige Herrscher, sowie ikonographische und epigraphische Informationen zu über 3000 Skulpturstücken.

3D-Modelle von Skulptur und Architektur, die 2013 mit airborne und terrestrischem Laserscanning und Fotogrammetrie und auch CAD angefertigt wurden, werden jetzt segmentiert, mit ihren archäologischen Attributen und Meta-daten verlinkt und als Testobjekte für das 3D-Dokumentationssystem verwendet.



Abb. 8: Scannen der Ruinen des Tempels 18, Copan (KAAK Fabio Remondino, FBK Trento)



Abb. 10: Vorläufige Modelle des Laserscans von Tempel 18 (KAAK: Fabio Remondino)

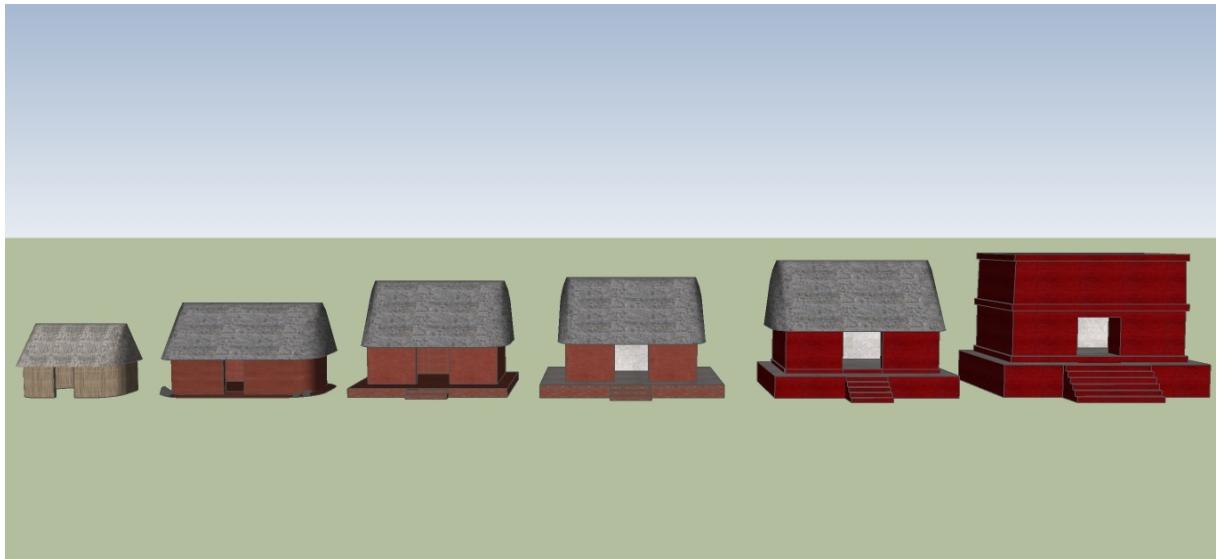


Abb. 11: CAD Modelle von verschiedenen Gebäudetypen in Copan (Heather Richards-Rissetto)

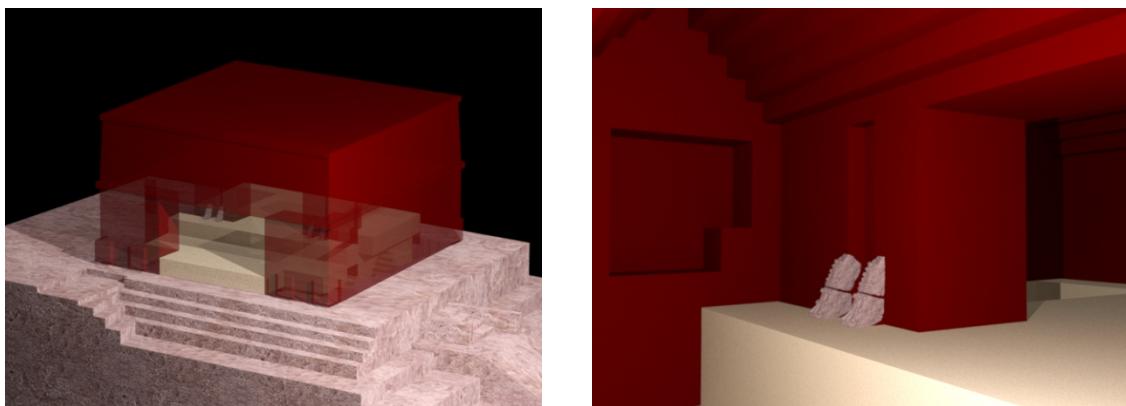


Abb. 12: Erste Versuche der Rekonstruktion des Tempels 18 in Copan (KAAK: Mike Lyons)

Derzeit wird an einem computergrafischen 3D-Modell von Tempel 18 in Copan gearbeitet, welches Wirklichkeits-basierende Modelle von Bauschmuckelementen mit virtuellen Rekonstruktionen oder Simulationen kombiniert. Zudem wurden 2013 LIDAR Daten von einer 24km² großen Fläche des Copan-Tales gesammelt und werden bis Ende des Jahres prozessiert.



Abb. 13: Erste Ergebnisse der LiDAR Daten von Copan (KAAK)

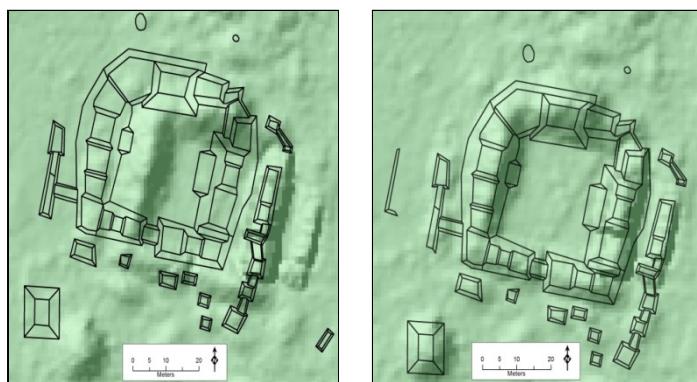


Abb. 14: Vergleich zwischen LiDAR Daten und Karten von Copan aus den 80er Jahren (Richards-Rissetto)

2014 wird das Teilprojekt Archäologie in Copan und Museumssammlungen weitere 3D Daten aufnehmen, strukturieren und in die Datenbank eingeben. Die ersten Analysen im online System werden durchgeführt und geprüft.

Teilprojekt Geoinformatik

Von 2012 bis 2013 prüfte das Teilprojekt Geoinformatik neue OpenSource Software-Optionen, die auf WebGL basieren, um eine neue Plattform aufzubauen und entschied sich für Three.js – eine JavaScript-Bibliothek in Kombination mit WebGL. Diese Kombination dient als Framework für die Entwicklung der webbasierten 3D Umgebung.

Die Geoinformatiker entwickelten auch einen Geobrowser (interaktive Landkarte, die verschiedene Ebenen von Daten visualisieren und abfragbar machen kann) auf Grundlage von Geomajas, einem Open-Source-GIS-Framework für das Web, welches der Anzeige und Bearbeitung der 2D- und 2,5D-Geodaten dient, und integrierte dazu ein „3D Single Object Viewer“ um hochauflöste 3D Modelle online zu visualisieren und zu analysieren.

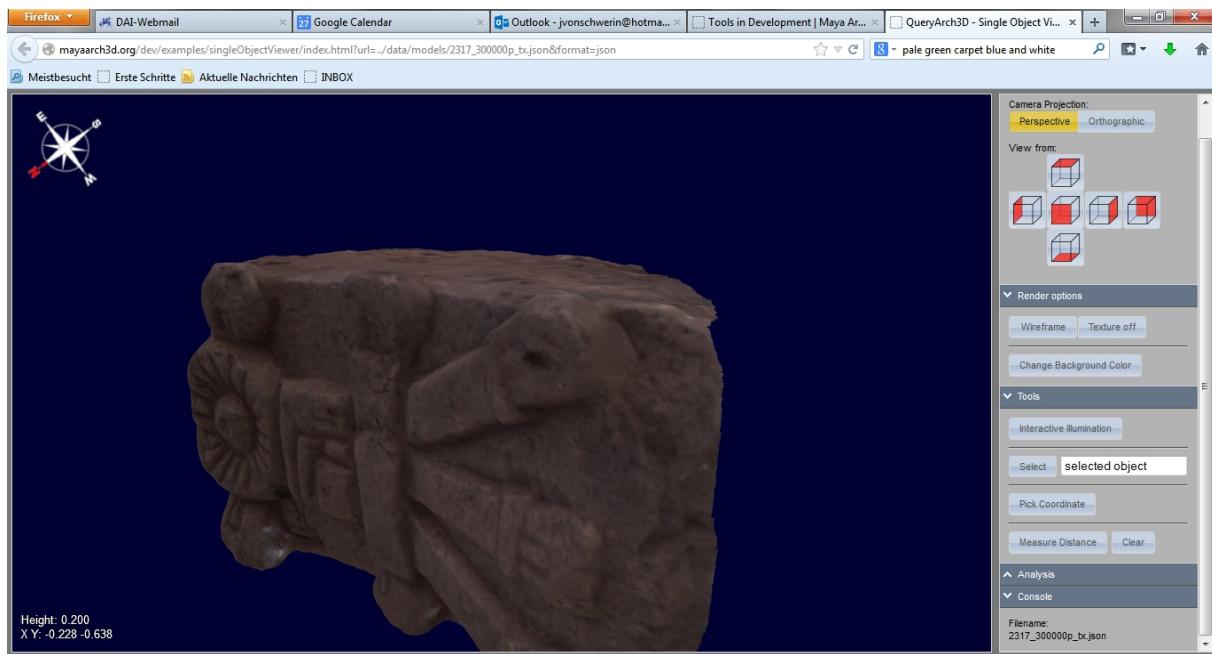


Abb. 15: Single-Object 3D Viewer

Bis Ende 2013 wird eine geeignete Schnittstelle für die Kommunikation zwischen der iDAIfield-Datenbank und der Datenbank für die 3D Objekte (Postgresql/PostGIS) entwickelt, damit Attribute und Meta-Daten von den (oft digitalen) archäologischen Objekten im Geobrowser und 3D Viewer visualisiert werden können. Ab Januar 2014 können in diesen zwei Modulen schon zweidimensionale archäologische Daten unterschiedlichster Art und Auflösung gespeichert, dargestellt und abgefragt werden. Hierbei wurde ein System zur Verwaltung der unterschiedlichen Nutzergruppen entwickelt, so dass die Daten gleichzeitig von mehreren Forschergruppen mit verschiedenen Niveaus von Nutzerrechten genutzt werden können.

Im kommenden Jahr werden die Geoinformatiker zudem die 3D-Umgebung für das System entwickeln, die raumbezogenen Abfragemöglichkeiten und GIS-Funktionen konstruieren, und eine Transparenzfunktion entwerfen, um 3D-Rekonstruktionen von den realitätsbasierten Modellen abzuhaben. Im- und Exportverfahren der unterschiedlichen 3D Datenformate werden auch unter Berücksichtigung internationaler Standards entwickelt.

Förderung: Bundesministerium für Bildung und Forschung, Gerda Henkel Stiftung

Leitung des Projekts: M. Reindel

Mitarbeiter und Mitarbeiterinnen:

Teilprojekt Archäologie: Wissenschaftliche Mitarbeiterin: J. von Schwerin; studentische Mitarbeiter: F. Fecher, M. Lyons, J. Meyer, L. Stelson,

Teilprojekt Geoinformatik: M. Auer, N. Billen, L. Loos, A. Zippf

Kooperationspartner: Deutsches Archaeologisches Institut, Kommission für Aufsereuropäische Archäologie, Geographisches Institut, Lehrstuhl Geoinformatik, und das Interdisziplinäre Zentrum für Wissenschaftliches Rechnen (IWR) der Universität Heidelberg; IT Abteilung, Deutsches Archäologisches Institut, Berlin; 3D Optical Metrology Group, Bruno Kessler Foundation (FBK), Trient, Italien; Honduranisches Institut für Archäologie und Denkmalpflege (IHAAH), Middlebury College, Vermont, USA.

Projektveröffentlichungen 2012-2013:

Von Schwerin, J., H. Richards-Rissetto, F. Remondino, and G. Agugiaro

2013 *The MayaArch3D Project: A 3D WebGIS for Analyzing Ancient Maya Architecture and Landscapes at Copan, Honduras. Literary and Linguistic Computing 2013*, doi: 10.1093/llc/fqt059. Oxford University Press.

Billen, N., Loos, L., Auer, M., Zipf, A., Richards-Rissetto, H., Reindel, M. & von Schwerin, J.

2013 Development of a 4D-webgis for archaeological research, Workshop on integrating 4D, GIS and cultural heritage, 16th AGILE 2013, Leuven.

Auer, M., Loos, L., Billen, N., Zipf, A., von Schwerin, J., Reindel, M. & Richards-Rissetto, H.

2013 *MayaArch3D - A web based 3D geoinformation system to analyse the archaeology of Copán, Honduras.- Poster at BMBF Kick-Off Workshop of the eHumanities joint projects 8.-9. April 2013 Leipzig*

Projektbericht: Wie arbeiten Digital Humanists für die Geisteswissenschaften?

Autoren: Marcel Schaeben, Bernhard Strecker, Gerrit Weber

Projektbetreuung: Prof. Dr. Manfred Thaller

Gegenstand: *Studiengänge in den Digital Humanities sollen Absolventen hervorbringen, die technische Lösungen aus dem Feld der Digital Humanities auch für GeisteswissenschaftlerInnen erstellen, die selbst dazu nicht in der Lage sind. Diese Form der Zusammenarbeit gilt auch als Desiderat in der gegenwärtigen Diskussion um den Einsatz der Neuen Medien im Umgang mit dem kulturellen Erbe[1]; trotzdem ist sie in der Praxis keineswegs unproblematisch. Wie sie in Form eines gemeinsamen Projekts am Ende der Ausbildung zu einem MA in den Digital Humanities erfahren wird, wird im Folgenden in einem Bericht der dafür verantwortlichen Projektgruppe am Beispiel einer konkreten mobilen Anwendung vorgestellt.*

[1] <http://www.creative-heritage.eu/>



Hintergrund

Anlässlich des bevorstehenden 450. Shakespeare-Jubiläums plant das Institut für Medienkultur und Theater der Universität zu Köln zusammen mit dem Museum für Angewandte Kunst Köln (MAKK) eine gemeinsame Ausstellung. Das Institut für Medienkultur und Theater greift dabei auf die umfangreichen Bestände der ihm angegliederten Theaterwissenschaftlichen Sammlung auf Schloss Wahn zurück.

Im Zuge dessen kam die Idee auf, in Kooperation mit dem Institut für Historisch-Kulturwissenschaftliche Informationsverarbeitung, an dem die Kölner Medieninformatiker ausgebildet werden, eine „Location-Based Gaming-App“ zu entwickeln. Die App wird Shakespeare und Kölsche Kultur in Symbiose vereinen. Die Spieler werden die Kölner Altstadt mithilfe von Mobile Devices und einer Kartenansicht erkunden (siehe Abb. 1). Auf der Karte werden verschiedene für die Spiele interessante kulturelle Orte wie Denkmäler, Sehenswürdigkeiten und typisch Kölsche Ecken ausgezeichnet. Sobald sich ein Spieler einem dieser Orte nähert (GPS), hat er die Möglichkeit kleine Spiele und narrative Versatzstücke zu starten. So erfährt er spielerisch etwas über Shakespeare und Köln.

Abb. 1: Mock-Up des Hauptbildschirms

Dabei sind diese Elemente in sogenannten Quests eingebettet (siehe Abb. 2.). Hier erlebt der Spieler kurze Geschichten, verwoben aus den benannten Elementen, die mit jeweils ortsspezifischen Hintergründen versehen sind. All diese Ressourcen stammen dabei aus der Zusammenarbeit mit der Theaterwissenschaftlichen Sammlung und wurden entsprechend der jeweiligen Anwendungsumgebungen aufbereitet. Mittels der eingebundenen Minigames wird neben dem Spielspaß auch die Immersion des Spielers gesteigert und eine lebhafte Einbindung in die Welten gefördert.

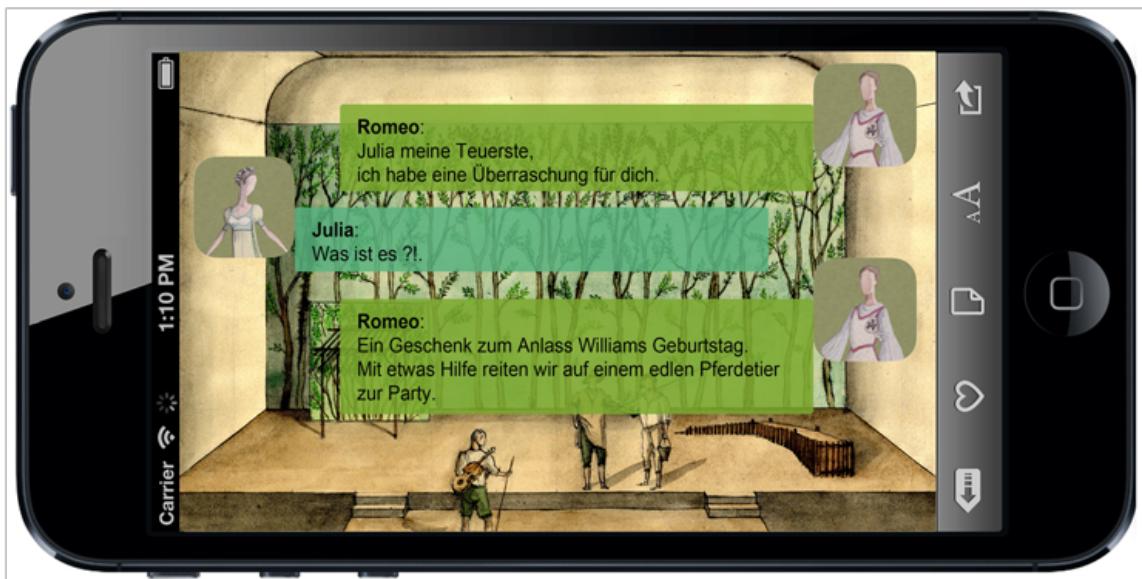


Abb. 2: Quest-Mock-Up

Ein vergleichbares, bereits realisiertes Projekt ist die *Zeitfenster-App*¹ stuttgarter Studenten, die allerdings mittels *Augmented Reality* Orte erkennt um Informationssysteme bereitzustellen.

Projektstruktur und Organisation

Alle eingeschriebene Medieninformatik studieren im Hauptfach Medienkulturwissenschaften und verfügen somit über ausgeprägte informationstechnologische wie geisteswissenschaftliche Kompetenzen. Angesprochen waren aber nicht nur Medieninformatiker, sondern auch Studierende, die andere Nebenfächer studieren (z. B. Anglistik, Medienmanagement, etc.). Die Verbundenheit durch die Medienkulturwissenschaft bietet eine gute Grundlage für die Zusammenarbeit, da alle Teilnehmer einen geisteswissenschaftlichen Kontext aufweisen.

¹ Vgl. <http://www.zeitfenster-app.de>; sowie http://www.creative-heritage.eu/4812.html?&tx_ttnews%5Btt_news%5D=6626&cHash=3b533def8fc18698ce7b89d589568f54 (31.12.2013).

Projektstruktur und Organisation wurden zu Beginn des Projektes nur grob festgelegt. Naheliegend war, dass sich die Medieninformatiker tendenziell eher den technischen Aspekten widmen, während die anderen Projektteilnehmer die Spielidee und -inhalte ausarbeiten. Die Rollen sind hierbei jedoch nicht besonders strikt angelegt, wodurch die entstehende Gruppendynamik optimal ausgenutzt wird. Durch diese Flexibilität wird eine effiziente Ressourcennutzung erreicht und Teilnehmer intervenieren - falls nötig - in anderen Entwicklungsprozessen. Werden Bereiche - vom permanent im Hintergrund ablaufenden Qualitätsmanagement - identifiziert, in denen es zu Problemen kommen könnte, werden diese personell verstärkt und neue Schnittstellen zu anderen spezialisierten Einheiten hergestellt.

Im Laufe des Projektes hat sich anhand der Interessen und Fähigkeiten der Projektteilnehmer sowie den jeweils aktuellen Projektanforderungen eine Aufteilung in drei Arbeitsgruppen herausgebildet: eine Narrations-, eine Layout- und eine Informatikergruppe. Die "Narrativler" nutzen ihre Expertise in den Theater- und Literaturwissenschaften, um einen groben Handlungsstrang sowie die einzelnen narrativen Elemente im Spiel kreativ auszuarbeiten. Die Layoutgruppe befasst sich mit den Konzeptarbeiten zur graphischen Benutzeroberfläche der App und setzt diese um, so dass sich in Verbindung mit dem locker-humorvollen Stil des Narrativs ein konsistentes künstlerisches Gesamtbild ergibt. Die Informatiker sind in erster Linie für die technische Umsetzung des Spiels zuständig, jedoch durch ihre Beraterrolle auch aktiv an der Entwicklung des Spielkonzeptes beteiligt. An der Schnittstelle der drei Gruppen steht ein eigens dafür eingesetzter Kommunikationsmanager, der steht vermittelnd und organisierend operiert.

Zwecks effizienter Koordination und Abstimmung wurde ein Wiki-System installiert, an dem alle Kursteilnehmer rege partizipieren. Es unterstützt zum einen den telemedialen Austausch über Problematiken (Issues) und aktuelle Fortschritte (commits) sowie einen – auch für Nicht-Informatiker – geregelten, leicht verständlichen Informationsfluss. Sogenannte Feedbackschleifen sorgen für kontinuierlichen Input der anderen Untergruppen und unterstützen die Entwickler bei der zielgerichteten und anwendungsorientierten Umsetzung. Die Entwicklung eines eigenen CMS-Systms ermöglicht komplexe narrative Inhalte übersichtlich und an einem gemeinsamen Ort strukturiert zu verwalten. Über ein einfach zu bedienendes Interface ist es allen Teammitgliedern möglich, kollaborativ Inhalte einzupflegen.

Projektablauf und aktuelle Entwicklung

Zur Realisierung des Projektes wurde ein grob strukturierter, zweisemestriger Kurs mit Aus- und Einstiegsmöglichkeit zur Halbzeit eingerichtet. Im ersten Semester erfolgte die Planung und Entwicklung des Spielkonzepts, Location-Scouting, Bestandsanalyse, Überprüfung technischer Umsetzbarkeit sowie die Verhandlung narrativer Dynamiken. Auch wenn der Schwerpunkt in dieser Phase auf der narrativen und konzeptuellen Ebene lag, waren die Informatiker von Beginn an vollwertige Kursteilnehmer.

Im Verlauf wurden diverse Konzepte erarbeitet, die einem internen Wettbewerb standhalten mussten, an dessen Ende die Ausarbeitung des jetzigen Narrativs stand. Bereits zu diesem frühen Zeitpunkt gaben die Informatiker der Gruppe Hilfestellung in Bezug auf die technischen Möglichkeiten und erstellen erste visuelle Repräsentationen (Mock-Ups). So war gewährleistet, dass alle Gruppenmitglieder auch weiterhin gemeinsam an einem maximal greifbaren Gegenstand mitwirken. Infolge gemeinsamer

Ortsbegehungen der Kölner Altstadt und der Recherche komplementärer Shakespearestücke nahm das Spiel sukzessive Gestalt an.

Zu Beginn des zweiten Semesters wurden neu hinzugestoßene Teilnehmer integriert sowie die technische Realisierung fokussiert, wodurch die Teilhabe der Medieninformatiker gewichtiger, die Arbeit der parallelen Gruppen aber nicht obsolet wurde.

Mit zunehmender Komplexität des Projektes zeigte sich, dass die Anforderungen an die gruppenübergreifenden Kommunikationsprozesse stark ansteigen. Zur Abfederung etwaiger negativer Auswirkungen wurde daher der Posten des Kommunikationsmanagers geschaffen, der ausschließlich für die Sicherstellung der Verständigung zwischen Informatikern und Narration zuständig ist sowie wöchentliche Treffen organisiert und moderiert. Dies hat sich als außerordentlich sinnvolle organisatorische Maßnahme erwiesen, welche die projektinterne Kommunikation optimiert und alle Beteiligten erheblich entlastet hat. Außerdem wurde nun das Layout-Team gegründet, das für alle Designaspekte zuständig ist, zwischen Narrationsgruppe und Informatikern vermittelt und das einheitliche Konzept und den Look der App weiter voranbringt.

Technische Umsetzung

Zu Beginn des Projektes wurde entschieden, dass eine möglichst große Zahl an Mobile Devices unterstützt werden soll. Als zentrale Plattformen wurden Android ab Version 4.0 sowie iOS ab Version 6 identifiziert, womit knapp 85% der verbreiteten Smartphones und Tablets abgedeckt werden (Stand Ende 2013)². Aufgrund der stark unterschiedlichen Programmierkonzepte und Schnittstellen beider Plattformen existieren separate Entwicklergruppen, die sich ausschließlich auf eine Plattform konzentrieren und deren native Möglichkeiten und Besonderheiten optimal nutzen können. Dennoch soll für alle Endbenutzer plattformübergreifend ein einheitliches Erscheinungsbild der späteren App gewährleistet sein. In regelmäßigen gemeinsamen Treffen beider Plattformgruppen, in denen bei Bedarf auch ein Mitglied der Layout-Gruppe anwesend ist, wird dies sichergestellt. Der Entwicklungsprozess ist an das agile Projektmanagement angelehnt, so dass Meilensteine und Deadlines jeweils flexibel an den aktuellen Entwicklungsstand und die sich dadurch ergebenden Anforderungen angepasst werden. So war von Beginn an der zeitliche Rahmen klar abgesteckt und eine Prioritätenliste festgelegt worden, nach der bestimmte Key-Features unbedingt integriert werden müssen, optionale Elemente aber nicht von vorn herein ausgeschlossen werden - je nach Auslastung der Entwickler kann daran gearbeitet werden.³ Augenmerk liegt hier aber stets auf der zeitigen Fertigstellung, wie der vorherigen ausgiebigen Qualitätskontrolle. Dazu werden frühzeitig Feldtests angesetzt, die mit Alphaversionen der App stattfinden und an der zahlreiche Teilnehmer und ein Pool von Außenstehenden, mit multiperspektivischem Blick auf Technik und Narration, teilnehmen. So wird sichergestellt, dass das Spiel in Gänze auch für "Uneingeweihte" Sinn ergibt, eine zweckmäßige Usability der GUI gegeben ist und letztlich der Spielspaß garantiert wird.

² Quellen:

<http://www.golem.de/news/android-verbreitung-jelly-bean-liegt-bei-45-1-prozent-zahlparameter-geaendert-1309-101418.html>

<http://www.areamobile.de/news/25931-apple-ios-7-ist-auf-74-prozent-der-ios-geraete-installiert>

<http://www.netmarketshare.com/> (Mobile/Tablet Operating System Market Share, Realtime Web Analytic)

³ vgl. auch <http://t3n.de/magazin/praxisbericht-scrum-kanban-scrumbuts-agiles-232822/>

Um diese Prozesse zu optimieren und zu vereinfachen, wurde ein eigenes CMS entwickelt, welches zur Verwaltung der Spielinhalte dient. Zudem fungiert es als gemeinsame Schnittstelle für die Narration, die Spielabläufe und Content einpflegt und dem Layout, welches Grafiken für die Oberfläche einbettet. Als weitere Funktion wurde der JSON-basierte Abgleich der App mit den Serverdaten implementiert, wodurch stehts aktualisierte Objekte geladen werden können, ohne den Clienten aktualisieren zu müssen. Positive Efekte sind zudem die einfachere Verwaltung des Contents, der stehts unabhängig bleibt, sowie eine stehts aktuelle Analyse technischer Schwierigkeiten.

Ausblick und Zwischenfazit

Geisteswissenschaftler und Informatiker sind in zwei unterschiedlichen Welten zu Hause. Die oben dargestellten Besonderheiten des Projekts – ein eigens eingesetzter Kommunikationsmanager, agiles Projektmanagement sowie das CMS als zentrale Arbeitsumgebung – münzen darauf, deren Zusammenarbeit so produktiv wie möglich zu gestalten und die Interdisziplinarität des Projektteams als Stärke und Chance zu nutzen. Dazu müssen Unterschiede in der Arbeitsweise sowie ganz unterschiedliche Grade an technischen Kenntnissen der Projektteilnehmer von Anfang an berücksichtigt werden. Diese Zusammenarbeit läuft selbstverständlich nicht immer reibungslos. Doch Ziel des Projektes ist es ja auch, neben der Fertigstellung des eigentlichen Spiels zu untersuchen, wie eine Kooperation zwischen Geisteswissenschaften und (Medien-)Informatikern, mit dem Ziel, neue Rezeptions- und Popularisierungsmöglichkeiten für kulturelle Inhalte zu erschließen, optimal ablaufen kann. Es gilt diesen Prozess zu untersuchen, daraus Lehren zu ziehen und diese als Impulse für zukünftige Kooperationsprojekte dieser Art zu nutzen. Gerade im Kooperationsaspekt steht unsere inspirierende Zusammenarbeit exemplarisch für die digitalen Geisteswissenschaften.

The History of Europe App - A pipeline for Humanist-Machine Interaction in the Digital Humanities

CUbRIK and the History of Europe App

The integration of human expertise and machine computation enables a new class of applications with significant potential for the digital humanities. So far this potential remains largely untapped due to the severe requirements of such projects: The implementation and integration of advanced algorithms requires specialized know-how and the final users from the humanities are challenged with defining unprecedented tasks for methods which haven't emerged yet. The FP-7-funded research project CUbRIK (www.cubricproject.eu) implements and integrates research in computer science, the design of human-computation tasks, data visualization, social engineering and the humanities.

In the proposed presentation we would like to showcase one of CUbRIK's case studies, the demo of the History of Europe application. The application introduces an effective interface to access collections of historical sources and to discover links among and entities within them. Upon completion CUbRIK will offer an innovative approach to human-enhanced time-aware multimedia search by synthesizing research in computer science, crowdsourcing and gamification. We will conclude the presentation with an outlook on the future development of the application.

Humanist-machine interaction

The History of Europe (HoE) application is based on a curated collection of more than 3000 images, representing the main events and actors in the history of the European integration. The collection is curated and hosted by the Centre Virtuel de la Connaissance sur l'Europe (CVCE). In a first step, an image indexation pipeline identifies the location of individual faces in the photographs. The location of these faces is verified by a crowd of "click-workers" with no specific training who evaluate for each recognized face if the depicted image shows a human face or not. Following the face verification process, an automatic face recognition process is triggered that associates each of the now verified faces with a list of ten possible identities. This list of candidates is then disseminated for example through Twitter to a crowd of experts that vote and comment for their preferred identity.

Besides the identities of the different persons, all information that is associated to an image, such as the time or the place where the image was taken as well as contextual information about associated historical events can be reviewed by expert users and delegated to a crowd of domain experts for review.

Data aggregation, visualisation and analysis

Building on the computed co-occurrence of persons in images a social graph is constructed that connects them with each other. Connections gain in strength the more often persons appear together in an image. Finally the result of this process is depicted in a visualization of the social graph with a set of analytical tools.

The social graph in the History of Europe App aims at representing and visualizing dependencies between historically relevant persons in the context of European integration. Thereby the weight of the (social) links between person entities relies on their co-occurrence in historic photographs as identified by the aforementioned image indexation process. The more frequently two persons appear in different photographs, the stronger the link between the corresponding entities in the graph.

Users can interact with the History of Europe social graph in different ways, e.g. a click on a node results on an ego-graph of the selected person and clicking on an edge displays documents that relate to both selected relationship. As the documents stored in the collection very often come with a date of creation, the graph can be filtered by date with the timeline, displaying only the connections of documents created within this timespan. This timeline also shows the amount of photos per date that are contained in the collection. Another filtering option is the number of connecting documents, which allows the visualization of those relationships that are only included in an interval of a minimum and maximum number of documents. This feature is useful to highlight highest co-occurrences. Finally, the number of appearances of a person in the processed collection lets us identify people who appear particularly often in any given time frame.

Crowd discussion and a new approach to the representation of truth in digital research tools

Another challenge for the HoE app and the domain of the Digital Humanities in general is the conception of truth, which differs significantly e.g. to the conceptions of truth in Computer Science. Computer Scientists can rely on a stable foundation of what is true: Any experiment can be replicated and measured precisely. In the humanities the concept of truth is far more complex: It is based on the insight, that there is no neutral or objective way to study human environments. The way, in which questions are asked, how data is selected to answer them, by what means this data is analyzed and finally the way in which the results of such analyses are communicated and received all challenge the idea of “one truth”.

In order to represent the discursive nature of truth in the humanities within HoE we make use of a community-driven tool for question answering, similar to stackoverflow.com. User have the opportunity to answer questions and thus benefit from the knowledge within the expert crowd. However, the system allows for more than one answer and offers its users the possibility to vote and answer up or down, thereby allowing more than one answer to enter in competition with each other whilst also maintaining the full spectre of the discussion.

Summary and outlook

The History of Europe application takes on the challenge to combine cutting edge research in the domains of computer science, the design of human-computation tasks, data visualization, social engineering and the humanities by identifying

synergies between the disciplines' strengths and by compensating for their weaknesses. We do this by building a pipeline which connects face recognition tools, data visualization and input from humans and creates an ongoing cycle of iteratively improved user input and machine output. The History of Europe application stands in line with a range of other online tools for historical research but introduces new social features as well as crowd sourcing from both click-workers and expert users which continuously improves the system. In the future we will expand the selection of sources to include digitized text documents as well as audio and video interviews from different archives.

Jahrestagung 2014 der Digital Humanities im deutschsprachigen Raum – Abstract

Beitragstyp: "Vortrag" (20 min Vortrag + 10 min Diskussion)

Titel: Frauenfragen um 1900 als Gegenstand kontroverser Kommunikation im Umkreis der >ersten< Frauenbewegung. Wie können digitale Ressourcen die Untersuchung und die Ergebnisdokumentation verbessern?

Projektgruppe

Dr. Kerstin Wolff

Stiftung Archiv der deutschen Frauenbewegung
Gottschalkstraße 57 – 34127 Kassel
Tel.: (0561) 9893-670 – Mail: wolff@addf-kassel.de

Kerstin Wolff ist Historikerin. Ihre Arbeitsschwerpunkte liegen in den Bereichen der historischen Frauenforschung und der Erforschung der >ersten< Frauenbewegung. Sie leitet die Forschungsabteilung bei der Stiftung Archiv der deutschen Frauenbewegung.

Dr. Alexander Geyken

Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)
Arbeitsstellen Digitales Wörterbuch der deutschen Sprache/Deutsches Textarchiv
Jägerstraße 22/23 – 10117 Berlin
Tel.: (030) 20370-390 – Mail: geyken@bbaw.de

Alexander Geyken ist Computerlinguist und Leiter der Arbeitsstellen des DWDS und des DTA an der BBAW. Seine Arbeitsschwerpunkte liegen in den Bereichen Sprachtechnologie, digitale Lexikographie, Korpuslinguistik und Korpustechnologie sowie Korpussstandards.

Prof. Dr. Thomas Gloning

Institut für Germanistik – JLU Gießen
Otto-Behaghel-Straße 10B – 35394 Gießen
Tel.: (0641) 99-29040 / -29041 (Sekr.) – Mail: thomas.gloning@uni-giessen.de

Thomas Gloning ist Sprachwissenschaftler (Germanistik). Fachliche Arbeitsschwerpunkte liegen u.a. in den Bereichen Kommunikationsanalyse, Textanalyse, Semantik, Wortschatzorganisation und Wortschatzdynamik, Geschichte von Kommunikationsformen und Texttypen. Darüber hinaus: Digitale Textkorpora und ihre Nutzung, digitale Infrastrukturen, Anwendbarkeit und Reichweite digitaler Ressourcen für geisteswissenschaftliche Fragestellungen.

Abstract

1. Arbeitstitel

Frauenfragen um 1900 als Gegenstand kontroverser Kommunikation im Umkreis der ›ersten‹ Frauenbewegung. Wie können digitale Ressourcen die Untersuchung und die Ergebnisdokumentation verbessern?

2. Zuordnung zu thematischen Schwerpunkten des *Call for papers*

- Wer bestimmt die [...] übergreifenden Forschungsagenden?
- Wie können Netze zur Darstellung und Präsentation geisteswissenschaftlicher Quellen und Ergebnisse genutzt werden? Wie können diese Ergebnisse in Forschungsinfrastrukturen integriert werden?
- Probleme des Markup (für spezifische fachliche Fragestellungen); Softwarewerkzeuge für die Geisteswissenschaften (hier u.a.. integrierte digitale Dokumentation)
- Disziplinspezifische Anwendungen digitaler Ressourcen; kuratorische Aspekte digitaler Verfahren (hier: thematische Spezialkorpora)

3. Beschreibung des Themas

Das im Folgenden beschriebene Thema steht für eine ganze Klasse von thematischen bzw. diskursorientierten Fragestellungen und die jeweils darauf bezogenen Textkorpora. Es ist damit ein paradigmatischer Fall für die Anforderungen, die geisteswissenschaftliche Projekte dieser Art an die Datenmanagementmethoden der Digital Humanities stellen.

3.1 Der Gegenstand: Texte und Diskurse der sog. ›ersten‹ Frauenbewegung

Die sog. ›erste‹ Frauenbewegung des 19. und beginnenden 20. Jahrhunderts war in ihrem Kern eine kommunikative Bewegung. Zentrale Streitpunkte und Forderungen in der Hochphase der ersten Frauenbewegung um 1900 betrafen vor allem die Bereiche politische Partizipation (Wahlrecht), Bildung (Mädchen Schulwesen, Hochschulzugang), Erwerbsarbeit (Zugang zu beruflichen Tätigkeiten; gerechter Lohn) und Sexualmoral (vor allem Prostitutionsdiskurs). Um solche Streitpunkte und Forderungen im öffentlichen Raum zu thematisieren, bedienten sich die Vertreterinnen der Frauenbewegung vor allem des geschriebenen Worts. In eigenen Zeitschriften, Petitionen, Flugschriften und Monographien legten sie Mo-

tive, Hoffnungen, Forderungen und Argumente dar und versuchten so, die Gesellschaft von der Notwendigkeit einer Veränderung der Geschlechterordnung zu überzeugen.

Während die *Themen* der >ersten< Frauenbewegung um 1900 in ihren sozial- und ideengeschichtlichen Grundzügen als gut erforscht gelten können, ist eine umfassende und detaillierte Untersuchung des *Sprachgebrauchs* in den öffentlichen Debatten um Frauenfragen des 19. und beginnenden 20. Jahrhunderts bisher nach wie vor ein Desiderat.

3.2 Fachliche Fragestellungen und Zielsetzungen

Eine systematische Analyse des Sprachgebrauchs und der kommunikativen Strukturen muss sich in erster Linie auf drei Teilfragen beziehen, die im Schnittpunkt von historischer Diskursanalyse, historischer Argumentationsforschung und historischer Lexikologie und Semantik zu verorten sind. Die Beantwortung dieser Fragen ist gleichzeitig ein Beitrag zur bislang nicht geschriebenen Kommunikationsgeschichte der >ersten< Frauenbewegung und ein Beitrag zu einer Geschichte des Sprachgebrauchs von Frauen.

- (1) Wie lassen sich einzelne **thematische und diskursive Stränge** rekonstruieren? Welche Einzeltexte gehören jeweils zu einem bestimmten diskursiven Strang? Welche intertextuellen Bezüge sind zwischen Einzeltexten und Teilen von Einzeltexten erkennbar? Mit welchen sprachlichen Verfahren etablieren die AutorInnen Bezüge zu anderen Texten, die entweder zur Stützung eigener Positionen oder aber als Beispiele für gegnerische Positionen angeführt werden?
- (2) Was sind zentrale **Thematisierungspraktiken** bei der Etablierung bestimmter Sichtweisen und welche **Argumentationsformen** werden gebraucht, um Sichtweisen und Forderungen zu stützen?
- (3) Wie lässt sich der **Wortgebrauch** dieser Texte in seiner spezifischen Funktionalität charakterisieren und dokumentieren? Teilfragen zum Wortgebrauch sind u.a.: Wie tragen unterschiedliche Formen des Wortgebrauchs dazu bei, Sichtweisen auf Geschlechterverhältnisse zu konstituieren und neue Forderungen zu stützen, durchzusetzen oder ihre Umsetzung zu verhindern? Welche Funktion haben unterschiedliche Wortschatzeinheiten bei der Organisation dieser Texte? Wie unterscheidet sich das lexikalische Profil von Texten aus dem Frauenfragen-Diskurs im Umkreis der ersten Frauenbewegung von Texten aus anderen Domänen?

3.3 Die Rolle digitaler Werkzeuge und Verfahren für die fachlichen Zielsetzungen

Digitale Werkzeuge und Verfahren spielen eine wesentliche Rolle in zwei unterschiedlichen Bereichen dieser Forschungsarbeit: (i) bei der Analyse und Bearbeitung der einzelnen Fragestellungen; (ii) bei der integrierten open-access-Präsentation der Quellentexte und der Darstellung der Analyse-Ergebnisse.

(i) Digitale Werkzeuge und Verfahren und ihre Rolle bei der Analyse

Im Rahmen unserer Vorarbeiten (siehe 3.4) haben wir erste Quellentexte als Volltexte erfasst, in standardkonformer Weise gemäß den Richtlinien der Text Encoding Initiative aufbereitet (genauer: gemäß dem DTA-Basisformat, welches Best Practice Format für die Repräsentation historischer Texte im Infrastrukturprojekt CLARIN-D ist) und im DTAQ-Bereich des Deutschen Textarchivs an der Berlin-Brandenburgischen Akademie der Wissenschaften, einem CLARIN-D-Partner, der den Zielen der Standardisierung, der Nachhaltigkeit, der Interoperabilität und der freien Nachnutzbarkeit verpflichtet ist, zur Verfügung gestellt. Digitale Ressourcen wurden im Bereich der Untersuchung bislang in erster Linie für die Zwecke der lexikalischen Analyse, in geringerem Umfang auch für die Zwecke der Analyse von Argumentationsformen und argumentativer Topoi eingesetzt.

(ii) Eine integrierte digitale Dokumentationsumgebung

Teil unserer Vorarbeiten waren auch konzeptionelle Überlegungen für einen neuartigen, integrierten Typ von digitaler Dokumentationsumgebung, die auf drei Säulen beruht:

- (a) **Darstellung von Untersuchungsergebnissen** zu Wortgebrauch und kommunikativen Verfahren/Thematisierungspraktiken in monographischer Form,
- (b) strukturierte digitale **Textcorpora** zu spezifischen Diskursbereichen und einzelnen Themensträngen,
- (c) erweiterbares digitales **lexikalisches System**, das systematisch auf die Untersuchungen (a) und die Textcorpora (b) bezogen ist. In diesem System werden zum einen die einzelnen Verwendungsweisen zentraler Ausdrücke lexikographisch beschrieben und auf den textuellen Gebrauch bezogen, zum anderen werden die einzelnen Bedeutungspositionen durch Deskriptoren markiert, so dass eine thematische, funktionale, gruppenspezifische usw. Erschließung des Wortgebrauchs ermöglicht wird.

Eine wesentliche Zielsetzung der lexikalischen Dokumentation ist es, dass die Resultate und Befunde anzubinden sind an laufende Wörterbuchprojekte, z.B. beim DWDS.

3.4 Eigene Vorarbeiten und bisherige Resultate

Zu den **Vorarbeiten** für den geplanten Vortrag gehören insbesondere:

- Ein vom Land Hessen (HMWK) gefördertes Pilot-Projekt (6/2011-12/2011);
- eine inzwischen abgeschlossene umfangreiche Gießener Magisterarbeit zu Thematisierungspraktiken und Wortgebrauch im Diskurs um die Mädchenschulreform;
- ein laufendes Promotionsvorhaben zu Wortgebrauch und Thematisierungspraktiken im Diskurs um das Frauenwahlrecht 1850-1918. Gegenstand dieser Arbeit ist auch die Zusammenstellung eines thematischen Korpus und eines digital nutzbaren und facettiert erschlossenen Glossars;
- Integration von digitalen Volltexten aus dem Themenbereich der Frauenfragen in das Deutsche Textarchiv im Rahmen eines CLARIN-D-Kurationsprojekts;

- Ausarbeitung von historischen, kommunikationsgeschichtlichen und sprachgeschichtlichen Fallstudien sowie von Prototypen der lexikalischen Dokumentation im Themenbereich des Vortrags.

Bisherige Resultate, über die wir in gebotener Kürze berichten werden bzw. können, sind:

- die historische, sprach- und kommunikationsgeschichtliche Verortung des Themas;
- die bisherigen fachlichen Resultate in den Bereichen der Wortschatzuntersuchung, der Argumentationsanalyse und der Untersuchung von Thematisierungspraktiken;
- die bisherigen Resultate und Erfahrungen im Bereich der Anwendung digitaler Ressourcen, v.a. in den Bereichen Volltextdigitalisierung, lexikalische Erschließung von Korpustexten, Organisation eines thematisch-diskursiv orientierten lexikalischen Informationssystems, Integration von Ergebnisdarstellung, Korpustexten und lexikalischer Dokumentation.
- Problemzonen bei der Nutzung digitaler Ressourcen für kommunikations-, diskurs- und sprachhistorischen Untersuchungen dieser Art.

4. Planungen zur Struktur des Vortrags

Der Vortrag soll drei wesentliche Teile aufweisen:

- (i) **Einführung** (erste Frauenbewegung; kommunikationsgeschichtliche Fragestellung; Frage nach der Rolle digitaler Ressourcen für Analyse und Ergebnisdarstellung);
- (ii) Gedrängter Überblick über die **bisherigen Resultate** und Erfahrungen;
- (iii) Im Mittelpunkt sollen dann die nächsten Schritte und offene bzw. diskussionswürdige Fragen stehen, die sich drei Bereichen des **Zusammenspiels von fachlich geprägter Forschung und der Anwendung von DH-Methoden** zuordnen lassen:
 - (a) Vorstellung des Konzepts einer integrierten open-access-**Dokumentationsumgebung**, die auf drei vernetzten digitalen Säulen beruht: monographische Darstellung von Analyseergebnissen, digitale Korpustexte, lexikalische Dokumentation.
 - (b) Verfahren der **Auszeichnung** und der **Auswertung** digitaler Korpustexte im Hinblick auf Thematisierungspraktiken und Argumentationsformen. Zu den Herausforderungen in diesem Bereich gehört, dass relevante Textteile von ganz unterschiedlicher Größe sein können und dass unterschiedliche Parameter der Textorganisation auch zu übereinanderliegenden Annotationsstrukturen führen.
 - (c) Digitale Unterstützung **lexikalisch-lexikologischer** Analysen von Korpustexten.

Zusammengefasst: Der Beitrag soll exemplarisch zeigen, welche Unterstützungspotentiale DH-Verfahren für fachliche Fragestellungen in den Bereichen Kommunikations- und Sprach- bzw. Sprachgebrauchsgeschichte aufweisen bzw. aufweisen müssen. Es sollen dabei auch die bisher erfahrenen Problemzonen und Reibungsverluste thematisiert werden.

„Losing My Religion“ – Einsatz einer Videoannotationsdatenbank in der kunstgeschichtlichen Analyse von Musikvideos

Am Institut für Europäische Kunstgeschichte der Universität Heidelberg hat sich seit 2012 mit der Untersuchung von Musikvideos (s.a. das aktuell laufende DFG-Projekt zur ästhetischen Umsetzung von Musikvideos im Kontext von Handhelds¹) ein Forschungsfeld herausgebildet, auf dem die Heidelberger Kunstgeschichte mit Prof. Dr. Henry Keazor eine internationale Führungsrolle einnimmt.

Ogleich das Thema des Musikvideos inzwischen eine verstärkte Aufmerksamkeit innerhalb von Disziplinen wie der Film-, Musik- und Medienwissenschaft sowie der Kunstgeschichte erfahren hat (vgl. dazu die Bibliografien der VertreterInnen dieser Disziplinen versammelnden Publikation von Keazor/Wübbena 2010, sowie die Tatsache, dass dem Musikvideo auf dem Kunsthistorikertag 2013 in Greifswald eine eigene Sektion gewidmet wurde), fehlt es in der wissenschaftlichen Bearbeitung des Themas derzeit noch an adäquaten Werkzeugen für die Analyse des Untersuchungsgegenstands und die Präsentation der Ergebnisse.

Zahlreiche Musikvideos weisen über ihren genuinen Zweck als Werbeträger hinaus einen ästhetischen Mehrwert auf, der eine eingehendere bildwissenschaftliche Betrachtung herausfordert und aufzeigt, dass sich in diesem Bereich eine eigene Kunstform entwickelt hat. Die künstlerischen Ausdrucksmöglichkeiten, die hier mit der flexiblen Verknüpfung von Text, Bild und Musik vorhanden sind, eröffnen Szenarien, in denen Text und Musik nicht nur einen kurzen Schwall von heterogenen Bildern zusammenhalten; vielmehr wird ein komplexer Diskurs auf drei Ebenen gestaltet. Das Musikvideo als visuelles Produkt referenziert dabei auch immer wieder auf „verwandte“ Medien (Kino- und Fernsehfilm), aber auch auf Werke der bildenden Kunst, der Zeitpolitik, anderer Kommunikations- und Unterhaltungsformen und auf sich selbst. Dieser Forschungsbereich stellt sich bisher durch diese enge Verbindung von Bild-, Text- und Musik/Audio-Ebenen (siehe hierzu auch die in Keazor/Wübbena 2005 [2011³] vorgestellte Analysestruktur zum Musikvideo) im Hinblick auf Informationsverwaltung und Visualisierung des komplexen Bezugsystems als Problemfall dar. Das Heidelberger Institut für Europäische Kunstgeschichte kann hier zwar auf eine der umfangreichsten Sammlungen digitalen Materials zu diesem Bereich zurückgreifen, jedoch fehlten bis dato die Werkzeuge und Visualisierungskonzepte, um die vielgliedrigen, intermedialen Bezüge zu erfassen und anschaulich darzustellen, welche für die wissenschaftliche Erschließung des Forschungsmaterials in diesem Feld unabdingbar sind.

¹ <http://portablemvs.net>

Im hier vorgeschlagenen Beitrag sollen nun am Beispiel des Musikvideos zum Titel „Losing My Religion“² der US-amerikanischen Rockband „R.E.M.“ – welches 1991 vom Regisseur Tarsem Singh umgesetzt wurde – die Möglichkeiten der „Video Annotation Database“, einer frei verfügbaren, quelloffenen Plattform zur kollaborativen, webbasierten Medienanalyse auf der Basis von „pan.do/ra – Open media archive“³, vorgestellt werden.

Das System hat sich in der Form bereits seit längerem im Exzellenzcluster „Asia and Europe in a Global Context“ der Universität Heidelberg (Cluster of Excellence) bewährt, wo es von der *Heidelberg Research Architecture* (HRA⁴) im Rahmen des Metadatenframeworks „Tamboti“⁵ genutzt und aktiv weiterentwickelt wird.

„Pan.do/ra“ ist eine Online-Applikation, die es – kurz gesagt – mehreren Nutzern gleichzeitig erlaubt, frei gewählte Abschnitte eines Videos mit unterschiedlichen Arten von Annotationen (Layer oder Spuren) zu versehen. Durch diese Annotationen ist es möglich, sowohl einzelne Sequenzen als auch sämtliche Szenen zu einer Thematik direkt über eine URL zu referenzieren. Jede Spur enthält zeitreferenzierte Annotationen eines bestimmten Typs und neben Text können auch georeferenzierte Ortsinformationen und Bilder hinzugefügt werden. Mit seiner integrierten Rechteverwaltung erlaubt es das System, Filme nur bestimmten Nutzergruppen zugänglich zu machen. So hat jüngst zum Beispiel eine Gruppe Studierender im Rahmen eines Geschichtsseminars die Applikation verwendet, um einen japanischen Propagandafilm aus den frühen 1930ern zu analysieren. Dabei wurden gezielt wiederkehrende Themen und Argumente im Film mit Schlagworten versehen, Texttafeln transkribiert und übersetzt, sowie Querverweise zu anderen Quellen hergestellt. Mit Hilfe der Videoannotationsdatenbank konnten die Studierenden die unterschiedlichen Argumentationsstränge im Film verorten und über die integrierte *Timeline* sichtbar machen, daneben ergab der Vergleich mit einer anderen (später entstandenen) Version des Filmes Unterschiede in der Narration.⁶

Die Einsatzmöglichkeiten des hier verwendeten Videoannotationssystems „pan.do/ra“ sind breit und generisch angelegt; so haben Max Stille und Jan Scholz beispielsweise islamische Predigt-videos analysiert und erste Ergebnisse davon wurden auf dem Deutschen Orientalistentag 2013 vorgestellt.⁷

Seit geraumer Zeit wird in Heidelberg an einer Integration in das Metadatenframework „Tamboti“ gearbeitet, womit Filmmetadaten gemeinsam mit bibliographischen Metadaten (in

² <https://vimeo.com/71613748>

³ <https://pan.do/ra>

⁴ <http://hra.uni-hd.de>

⁵ <http://tamboti.uni-hd.de>, <http://about-tamboti.uni-hd.de/>

⁶ Die Ergebnisse des Projekts sind unter <http://lytton-project.uni-hd.de> einzusehen.

⁷ <http://www.dot2013.de/programm/abstracts/panel-asthetik-und-oralitat/>

MODS XML), Bildmetadaten (in VRA Core 4 XML) und mit TEI ausgezeichneten Texten durchsuchbar gemacht werden sollen. Ziel der Integration ist es „pan.do/ra“ noch weiter für die stark interdisziplinäre Verbundforschung – wie sie z.B. am Exzellenzcluster „Asien und Europa“ betrieben wird – zu öffnen und auch über die Grenzen des Clusters hinaus in der Forschungsinfrastruktur zu verankern.

„Pan.do/ra“ kommt aufgrund seiner die verschiedenen Schichten audiovisuellen Materials visualisierenden und erschließenden Struktur dem verfolgten Forschungsansatz des Heidelberger Lehrstuhls für Neuere und Neueste Kunstgeschichte besonders entgegen. Diese Plattform bietet die Möglichkeit, das gelegentlich äußerst dichte Beziehungsgeflecht, welches das Musikvideo mit anderen Artefakten eingehen kann, zu erfassen und nachzuzeichnen. Nach dem Einspeisen kann das Musikvideo in der Folge konsultiert, analysiert und mit entsprechenden Notaten wie beispielsweise Transkriptionen, Beschreibungen, Schlagwörtern aber auch anderen Medien oder Querverweisen versehen werden. Auf diesem Wege kann der Medienbruch vermieden und das Bezugssystem direkt am Objekt verdeutlicht werden – im Falle des hier betrachteten Clips „Losing My Religion“ u.a. durch Referenzen zu Werken von Caravaggio, Andrei Tarkowski und Pierres et Gilles, aber auch zu Otto Lilenthal und Vaslav Nijinsky. Neben dieser Option, Informationen anderer Datenbanksysteme und Quellen zu referenzieren, erlaubt das webbasierte OpenSource-System den gemeinsamen Zugriff auf das Material, was somit eine ideale Basis für evolutionäres Erarbeiten von Ergebnissen in Forschergruppen und natürlich auch mit Studierenden bietet.

Im Rahmen des vorgeschlagenen Beitrags möchten wir anhand des oben genannten Musikvideo-beispiels die Möglichkeiten des genutzten Systems in Abgrenzung zu vergleichbaren Systemen vorstellen. Neben den Funktionen des *Frontends* wird auch kurz auf die technische Anbindung an ein bestehendes Authentifizierungssystem und den Datenaustausch mit anderen Systemen über Schnittstellen eingegangen. Hier steht insbesondere die Integration in das Metadatenframework „Tamboti“ im Vordergrund und es wird gezeigt, wie in einer kollaborativen Arbeitsumgebung Ressourcen medien- und kollektionsübergreifend mit wissenschaftlichen Annotationen versehen werden können.

Perspektivisch werden Optionen zur Weiterentwicklung aufgezeigt und dabei auch die Fragen nach Nachhaltigkeit und Pflege von Daten, Metadaten und Datenbanksystemen einzbezogen.

Hauptziel des Beitrags wäre auch die Verdeutlichung des Umstands, dass durch die Verzahnung der Informationstechnologie mit der Fachwissenschaft ein Mehrwert erreicht werden kann, der

zuvor durch die begrenzten analogen Möglichkeiten resp. auf anderem Wege nicht hätte erzielt werden können – ein geradezu idealtypisches Modell aus/auf dem Gebiet der „Digital Humanities“.

Publikationen der Einreicher zum Thema

Wübbena, Thorsten: „Video thrills the Radio Star“: Musikvideos: Geschichte, Themen, Analysen (als Autor, zusammen mit Henry Keazor), Bielefeld 2005 (3. Auflage: 2011).

Wübbena, Thorsten: Rewind – Play – Fast Forward: The Past, Present and Future of the Music Video (als Herausgeber zusammen mit Henry Keazor), Bielefeld 2010.

Wübbena, Thorsten: Imageb(u)ilder: Vergangenheit, Gegenwart und Zukunft des Videoclips (als Autor sowie, zusammen mit Thomas Mania und Henry Keazor, als Herausgeber), Münster 2011.

Wübbena, Thorsten: Zur ästhetischen Umsetzung von Musikvideos im Kontext von Handhelds (als Autor wie als Herausgeber zusammen mit Hans Giessen und Henry Keazor), Heidelberg 2012 (online verfügbar unter <http://archiv.ub.uni-heidelberg.de/ardok/volltexte/2012/1867/>).

Wübbena, Thorsten: Kapitel „Music Video“, in: *See this Sound. An Interdisciplinary Survey of Audovisual Culture*, Köln 2010, hrsg. von Dieter Daniels/Sandra Naumann und Jan Thoben, S. 223-233 (zusammen mit Henry Keazor; online verfügbar unter <http://see-this-sound.at/kompendium/abstract/44>).

Sun, Liying and Arnold, Matthias: “TS Tools. Designing a database of historical periodicals for research and teaching: The Chinese Women’s Magazines database.” In *Tijdschrift Voor Tijdschriftstudies* no. 33 (November 7, 2013): 73–78. doi:<http://bit.ly/15ndLtv>.

Arnold, Matthias, K. Berner, P. Gietz, K. Schultes, and R. Wenzlhuemer. “GeoTwain: Geospatial Analysis and Visualization for Researchers of Transculturality.” In *2009 5th IEEE International Conference on E-Science Workshops*, 175–179. Oxford, 2009.
doi:10.1109/ESCIW.2009.5407969.

—

Film with Oliver Seibt on Visual-Kei. Digital video, Filmportraits. Cluster „Asia and Europe“, 2012. <http://www.asia-europe.uni-heidelberg.de/en/research/heidelberg-research-architecture/hra-portal/detail/m/film-portrait-of-oliver-seibt.html>.

“Global Politics on Screen – A Japanese Film on the Lytton Commission in 1932” Projektwebsite, 2012. <http://lytton-project.uni-hd.de>.

Kurzprofile

Matthias Arnold hat Kunstgeschichte Ostasiens und Sinologie studiert und betreut die visuellen Ressourcen innerhalb der Heidelberg Research Architecture (HRA) am Exzellenzcluster „Asien und Europa“. Neben der Betreuung von Digitalisierungsprojekten am Cluster und dem angeschlossenen MediaLab gehören Konzeption und Umsetzung von visuellen Datenbanken sowie deren Einsatz in Forschung und Lehre zu seinem Interessens- und Arbeitsgebiet. Zu den größeren Datenbankprojekten gehören unter anderem „Chinese Women’s Magazines“ und die „Priya Paul Collection“. Aktuell arbeitet er an der Konzeption und Entwicklung von *Ziziphus*, eines auf dem VRA Core 4 XML Standard basierenden Editors für Bild-Metadaten, der Teil des Metadatenframeworks „Tamboti“ ist.

Eric Decker hat Politikwissenschaft Südasiens studiert und ist Koordinator der Heidelberg Research Architecture (HRA) am Exzellenzcluster „Asien und Europa“. Neben seiner Tätigkeit als Koordinator betreut er die Nutzer der Videoannotationsdatenbank „pan.do/ra“ und hat bereits zwei Seminare (Musikethnologie und Geschichte) betreut in denen das System eingesetzt wurde.

Thorsten Wübbena hat Kulturwissenschaften, Kunstgeschichte und Geschichte studiert. Seit 2000 wiss. Mitarbeiter im Kunstgeschichtlichen Institut der Goethe-Universität Frankfurt am Main. Von 2007 bis 2012 wiss. Leitung (gemeinsam mit Anna Schreurs und Carsten Blüm) im DFG-Projekt „Sandrart.net: Eine netzbasierte Forschungsplattform zur Kunst- und Kulturgeschichte des 17. Jahrhunderts“ (Universität Frankfurt, KHI Florenz). Seit April 2011 wiss. Mitarbeiter im DFG-Projekt „Zur ästhetischen Umsetzung von Musikvideos im Kontext von Handhelds“ (Universität des Saarlandes, Universität Heidelberg). Arbeitsschwerpunkte liegen im Bereich der Informationstechnologie in der kunstgeschichtlichen Forschung (Digitale Kunstgeschichte) sowie der Musikvideos.