
Der Gang durch die Domänen

zur Erfassung, Aufbereitung und Präsentation von Audiodaten im BMBF-Projekt „Freischütz Digital“

Benjamin W. Bohl · Thomas Prätzlich · Meinard Müller · Joachim Veit



Abb. 1 Exemplarische Darstellung von Archivinhalten – Noten- und Librettotexte (jeweils in Faksimile und Transkription) sowie Audioaufnahmen.

Das BMBF-Projekt *Freischütz Digital*¹ (FreiDi), widmet sich der paradigmatischen Konzeption und Umsetzung eines genuin digitalen Editionskonzepts am Beispiel von Carl Maria von Webers Oper *Der Freischütz*.

Die International Audio Laboratories Erlangen sind eine gemeinsame Einrichtung der Friedrich-Alexander-Universität Erlangen-Nürnberg und des Fraunhofer-Instituts für Integrierte Schaltungen IIS.

Benjamin W. Bohl und Joachim Veit
Musikwissenschaftliches Seminar Detmold/Paderborn
Gartenstr. 20
32756 Detmold
Tel.: +49 5231 975-876
E-Mail: thomas.praetzlich@audiolabs-erlangen.de

Thomas Prätzlich und Meinard Müller
International Audio Laboratories Erlangen
Am Wolfsmantel 33
91058 Erlangen – Tennenlohe
Tel.: +49 9131 85-20 520
E-Mail: thomas.praetzlich@audiolabs-erlangen.de

¹ <http://www.freischuetz-digital.de>

Leitgedanke ist Wierings *Multidimensional Model* [6], das im Sinne eines *kritischen Archivs* eine dichte Verknüpfung und Annotierung von Quellen unterschiedlicher medialer Ausprägung vorsieht (siehe Abb. 1). Ein für *FreiDi* wichtiger Aspekt ist in diesem Zusammenhang der erstmalige Einbezug von Audio-Daten in den Kontext einer Digitalen Edition (Detmold). Hierfür werden Algorithmen zur automatisierten Audiosegmentierung eingesetzt und entwickelt (Erlangen) [4, 5].

In dieser Kombination von Musikinformatik und Musikwissenschaft entstehen (für ausgewählte Aufnahmen des *Freischütz*) strukturell und inhaltlich reiche Metadaten, die die graphische, logische und akustische Domänen in Bezug setzen [1]. So werden zum Beispiel Pixelpositionen (graphische Domäne) mit Noteninformationen (logische Domäne) und Zeitpositionen (akustische Domäne) verknüpft. Die Codierungsrichtlinien der *Music Encoding Initiative*² (MEI) liefern hierfür einen umfassenden Rahmen.

Während für die Musikwissenschaft die Kombination von graphischer und akustischer Domäne eine perspektivische Weitung im Kontext einer Digitalen Edition ermöglicht [2], und damit etwa neue Möglichkeiten für die Rezeptions- und Interpretationsforschung bietet, kann auch die Musikinformatik von der engen Verknüpfung der logischen mit der akustisch-performativen Domäne profitieren. So können zum Beispiel die mit den Zeitpositionen verknüpften Notentextdaten zur Verbesserung der Ergebnisse von Quellentrennungsalgorithmen genutzt werden [3]. Hierbei ist das Ziel einzelne Instrumentenstimmen aus einer Audioaufnahme, die verschiedene gleichzeitig erklingende Stimmen enthält, abzutrennen.

Dieses Poster soll zunächst die Anforderungen der beiden Disziplinen an die Datenmodellierung und die jeweilige Umsetzung in *MEI* vorstellen. Schließlich sollen unter dem Gesichtspunkt der Präsentation und

² <http://www.music-encoding.org>

Abb. 2 *FreiDi-scoreFollowsAudio* Screenshot des Demonstrators zur synchronisierten Anzeige gerendeter Noten zur Audiowiedergabe. Der blaue Kasten hebt den aktuell erklingenden Takt automatisch hervor.

Abb. 3 Screenshot des Demonstrators zum Abspielen von Multitrack Aufnahmen. Die Mikrofon-Symbole sind entsprechend der Aufnahmesituation angeordnet – ein Klick darauf spielt das jeweilige Signal ab.

Nachnutzung der Daten die im Projekt entwickelten Demonstratoren und Web-Applikationen vorgestellt werden, etwa zum automatischen Blättern eines Notentexts zu einer laufenden Aufnahme (siehe Abb. 2), oder zum veranschaulichenden Abspielen von Multitrack Aufnahmen (siehe Abb. 3). Dabei gilt es auch der Frage nachzugehen, ob *MEI* für Echtzeit-Anwendungen sinnvoll einsetzbar ist.

Literatur

1. Babbitt, M.: The use of computers in musicological research. *Perspectives of New Music* **3**(2), 74–83 (1965)
2. Bohl, B., Kepper, J., Röwenstrunk, D.: Perspektiven digitaler Musikeditionen aus der Sicht des Edirrom-Projekts. In: *Die Tonkunst* **5**, 270–276 (2011)
3. Müller, M., Driedger, J., Ewert, S.: Notentext-informierte Quellentrennung für Musiksignale. In: *Proceedings of 43th GI Jahrestagung, 2928–2942*. Koblenz, Germany (2013)
4. Prätzlich, T., Müller, M.: Freischütz digital: A case study for reference-based audio segmentation of operas. In: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 589–594. Curitiba, Brazil (2013)
5. Prätzlich, T., Müller, M.: Frame-level audio segmentation for abridged musical works. In: *Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR)*. Taipei, Taiwan (2014)
6. Wiering, F.: Digital critical editions of music: A multidimensional model. In: *Modern Methods for Musicology*, 23–45 (2009)

Poster

***Big, complex, heterogeneous..* Laufende Projekte aus dem Arbeitsbereich *Big Data in den Geisteswissenschaften* in DARIAH-DE**

Stefan Pernes, Uni Würzburg

Der Begriff *Big Data* wird in den unterschiedlichsten Kontexten gebraucht, er umfasst unterschiedliche Größenordnungen und Strategien der Datenverarbeitung, und kann aufgrund dieser heuristischen Schwäche im besten Fall als *Buzzword*, mit Sicherheit aber nicht als trennscharfes Konzept bezeichnet werden. Zu Recht wird die Aufmerksamkeit zunehmend auf Fragen der Reliabilität und Validität der Verfahren gelenkt (Jordan 2014) und übergroße Heilsversprechen sowie Ankündigungen eines *Endes der Theorie* (Anderson 2008) kritisch hinterfragt. In geisteswissenschaftlichen Kontexten rückt diese allgemein geführte Diskussion jedoch in den Hintergrund. Hier gilt es, Textbestände in einer zuvor nicht da gewesenen Größe unter Berücksichtigung ihrer Vielschichtigkeit und Heterogenität zu verwalten und festzustellen, welchen Beitrag quantitative Methoden zu hermeneutischen Interpretationsverfahren leisten können. Diese Spezifika führen auch dazu, dass einige Voraussetzungen erst geschaffen werden müssen; so zum Beispiel das Training bestehender Verfahren der Sprachverarbeitung für literarische Textsorten, das Erstellen spezifischer Korpora und Vokabulare, oder die Verbesserung der Texterkennung von mittelalterlichen Handschriften. Das sind einige der Aufgaben, zu denen die *Use Cases* des DARIAH-DE Clusters *Big Data in den Geisteswissenschaften* einen Beitrag leisten und die im Folgenden vorgestellt werden sollen. Die *Use Cases* bearbeiten Fragestellungen aus Literaturwissenschaft, Philologie und Geschichte und werden jeweils in Kooperation eines fachwissenschaftlichen Partners und eines Partners aus dem Bereich der angewandten Informatik durchgeführt.

Narrative Techniken und Untergattungen im Deutschen Roman

Lehrstuhl für Computerphilologie, Uni Würzburg / Ubiquitous Knowledge Processing Lab, TU Darmstadt

Fachwissenschaftlicher Gegenstand des *Use Case* ist es, anhand quantitativer Verfahren die historische Entwicklung narrativer Techniken und in weiterer Folge auch die Entwicklung darauf aufbauender literarischer Kategorien nachzuvollziehen. Die Textgrundlage bildet ein Korpus 2000 deutschsprachiger Romane aus dem Zeitraum von 1500 bis 1930 und eine Sammlung von 200 französischen Kriminalromanen aus dem 19. und 20. Jahrhundert. Zum Einsatz kommen Verfahren zur automatischen Erkennung bestimmter Merkmale, wie zum Beispiel der Erkennung von Eigennamen oder von Passagen direkter Rede. Sämtliche Merkmale werden im Anschluss als *Features* zueinander in Bezug gesetzt, um die Texte

zu gruppieren und Gattungsbegriffe nachprüfen zu können. Parallel dazu werden Lernmaterialien erstellt, welche die dabei entwickelten Arbeitsabläufe in Form allgemeinverständlicher *Rezepte* zugänglich machen und interessierte ForscherInnen dazu ermächtigen sollen, *state of the art* Werkzeuge der Sprachverarbeitung für ihre jeweiligen Forschungsvorhaben einzusetzen und auf ihre jeweiligen Daten anzupassen.

Identifikation grenzübergreifender Lebensläufe in nationalen Biografien

Leibniz-Institut für Europäische Geschichte Mainz / Lehrstuhl für Medieninformatik, Uni Bamberg

Der *Use Case* erforscht die Verbindungen von individuellen historischen Lebensläufen und Internationalitätskriterien auf Grundlage von *Wikipedia* und mehreren europäischen Nationalbiografien. Dabei werden verschiedene Merkmale von Mobilität - wie zum Beispiel Geburts-, Wirkungs- und Sterbeorte, Tätigkeiten und verwandtschaftliche Beziehungen - miteinander korreliert und durch eine gezielte Erhebung sämtlicher Zusammenhänge mitunter Beobachtungen gemacht, die in den Geschichtswissenschaften noch nicht theoretisch erfasst sind. Die Datengrundlage umfasst strukturierte Daten sowie unstrukturierte Texte in mehreren Sprachen die miteinander verschränkt werden. Zusätzlich zum fachwissenschaftlichen Erkenntnisgewinn, stellt das Vorhaben eine quantitative Grundlage für kontrollierte Vokabulare in der Biografieforschung dar und zeigt auf, welche inhaltlichen und formalen Kategorien für *Semantic Web*-Ansätze in der Biografieforschung erforderlich und nutzbringend sein können.

Spuren der Zitation und Wiederverwendung im OpenMigne Korpus

Lehrstuhl für Digital Humanities, Uni Leipzig / Lehrstuhl für Medieninformatik, Uni Bamberg

Ausgehend von Editionen der Texte frühchristlicher Kirchenväter durch Jacques Paul Migne im 19. Jahrhundert entwickelt der *Use Case* Verfahren zur Erschließung vollständiger diachroner *Zitationsspuren*. Es handelt sich dabei um ein Feststellen chronologisch nachvollziehbarer Verläufe in einem Netzwerk von Zitationen, welches sich über ein gesamtes Korpus spannt. Textgrundlage bildet das *OpenMigne* Korpus, dessen Texte in griechischer und lateinischer Sprache einen Zeitraum vom Ursprung des Christentums bis in das 15. Jahrhundert abdecken. Die technische Umsetzung verläuft schrittweise: Beginnend mit der Erkennung von Zitationen in direkter Rede und in gleichsprachigen Ursprungstexten werden die Verfahren dahingehend erweitert, dass auch eine Erkennung von Paraphrasierungen und Zitationen in sprachlich heterogenen Korpora möglich wird. Die entwickelten Verfahren werden weiters für eine Anwendung über den *Use Case* hinaus aufbereitet und bereitgestellt.

Fazit

Anhand der vorgestellten Projekte wird ein Mal mehr deutlich, wie unterschiedlich die Voraussetzungen und Fragestellungen sein können, die unter dem Begriff *Big Data* verhandelt werden. Dabei tritt jedoch auch in den Vordergrund, was in diesem Feld die gemeinsamen, spezifisch geisteswissenschaftlichen Interessen sein können - ein methodologischer Austausch, von dem alle beteiligten Disziplinen profitieren.

Literatur

Jordan, Michael (2014): *Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts*. Interview by Lee Gomes, IEEE Spectrum. <http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts> (10.11.2014)

Anderson, Chris (2008): *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, Wired Magazine 16.07. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory (10.11.2014)

An End-To-End Integration of Automatic Annotations into CATMA

Thomas Bögel* Marco Petris† Jannik Strötgen* Michael Gertz*

* Institute of Computer Science, Heidelberg University
{boegel, stroetgen, gertz}@uni-hd.de

† Institute for German Studies, University of Hamburg
marco.petris@uni-hamburg.de

1 Introduction

Natural Language Processing offers solutions for predicting linguistic annotations at different levels of complexity. Thus, it seems obvious and – in general – a good idea to apply these methods to the Humanities in order to automate laborious manual annotations and to facilitate a deeper text analysis understanding. Apart from the purely technical aspect of developing suitable models, however, additional challenges for NLP in the Humanities arise: in order to be used as part of an analysis tool, humanists often desire justifications and explanations of automatic annotations. Just implementing a black-box approach, evaluating it intrinsically and returning the presumably best results to the user is not sufficient. In this paper, we suggest a transparent way of presenting the results of a NLP pipeline in a collaborative setting. This gives the user the possibility to judge the results directly within an already existing annotation interface and potentially use them for individual analysis tasks.

We will first present individual components that are combined with each other, namely the collaborative annotation tool CATMA and UIMA as a processing pipeline for Natural Language Processing. We will then show our end-to-end integration of UIMA into CATMA and its advantages.

2 CATMA integration

CATMA¹ is a flexible, collaborative annotation tool for literary scholars. So far, it integrates three functional and interactive modules, namely the tagger, the analyzer, and the visualizer. While the tagger module is a graphical interface to allow the easy

creation of manual annotations in texts using flexible tag sets (including feature structures, overlapping annotations, etc.), the analyzer component offers a wide range of possibilities to query a document collection or single documents, e.g., for frequently occurring patterns. Finally, the visualizer module can be used to explore a document collection, e.g., by generating distribution charts of the analysis results. In this paper, we present an extension to CATMA, which was developed in the context of the heureCLÉA project² - the integration of a UIMA-based text processing pipeline for the automatic creation of tag annotations created by natural language processing tools.

UIMA (Unstructured Information Management Architecture)³ is a wide-spread framework for developing and using natural language processing pipelines. One of its key characteristics is that it allows the easy combination of tools that have initially not been built to be used together. All UIMA components rely on the same data structure - the Common Analysis Structure (CAS) - there are three types of components: collection readers, analysis engines, and CAS consumers. The collection readers task is to access the source of the documents that are to be processed and to initialize a CAS object for each document. Then, the analysis engines perform linguistic processing of the data and stand-off add annotations to the CAS object. The subsequently called analysis engines can access the annotation results of the earlier components, i.e., they can perform more complex tasks. Finally, a CAS consumer performs the final processing of the CAS object.

¹Website: <http://www.catma.de/clea>

²<http://heureclea.de/>

³Website: <http://uima.apache.org/>

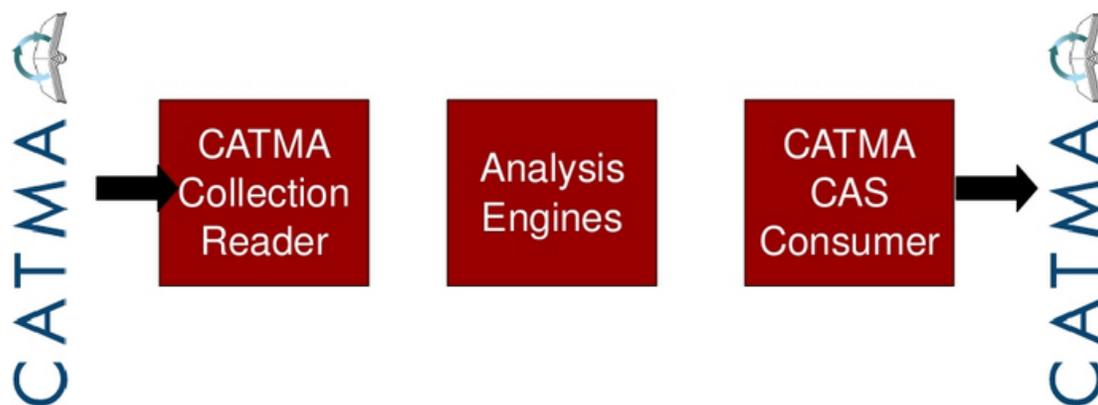


Figure 1: End-To-End architecture of combining the collaborative annotation platform CATMA with the automatic text processing pipeline UIMA.

In our case, the pipeline architecture is set up as depicted in Figure 1. The Collection Reader accesses the documents directly from CATMA and returns annotation information back to CATMA. However, the actual key feature of our development is that the user can directly access the automatic processing feature within the CATMA interface. That is, the user can select the types of annotations that shall be added to her document or document collection automatically. This significantly decreases the boundary for users not familiar with applying NLP tools for automatic processing of textual data, i.e., for typical CATMA users who are often literary scholars or students of the Humanities.

Nevertheless, our implementation is not a black box solution that only adds annotations that the user has to accept. In contrast, we are currently working on integrating a user feedback interface that will allow the initialization of user parameters based on the users feedback in the form of accepted or rejected annotations.

3 Research Workflow within CATMA

The advantage of a direct integration of UIMA into CATMA is best illustrated with an example: in order to analyse the temporal structure of documents (such as order phenomena), many linguistic aspects need to be taken into account. Temporal signals, e.g., calendrical, deictic or relational temporal expressions (Lahn and Meister, 2008), offer a hint for temporal phenomena of order. As manual annotation for these

basic linguistic phenomena is laborious, we are currently developing a machine learning system for predicting temporal signals. Figure 2 shows the possibility to create and directly inspect automatic annotations directly within the CATMA interface. With one click, the prediction of our NLP pipeline for temporal signals – or other annotations such as date and time expressions (Strötgen and Gertz, 2013) – can be shown. Note that the system output can easily be compared to any manual annotation as the type systems are completely independent. This flexibility allows scholars to focus on complex phenomena of the text with the possibility of automating simpler annotations. All automatic annotations are, however, non-obtrusive and completely changeable and reversible to give the choice of the level of automation to the user.

References

- Lahn, S. and J. C. Meister (2008). *Einführung in die Erzähltextanalyse*. Stuttgart: JB Metzler.
- Strötgen, J. and M. Gertz (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation* 47(2), 269–298.

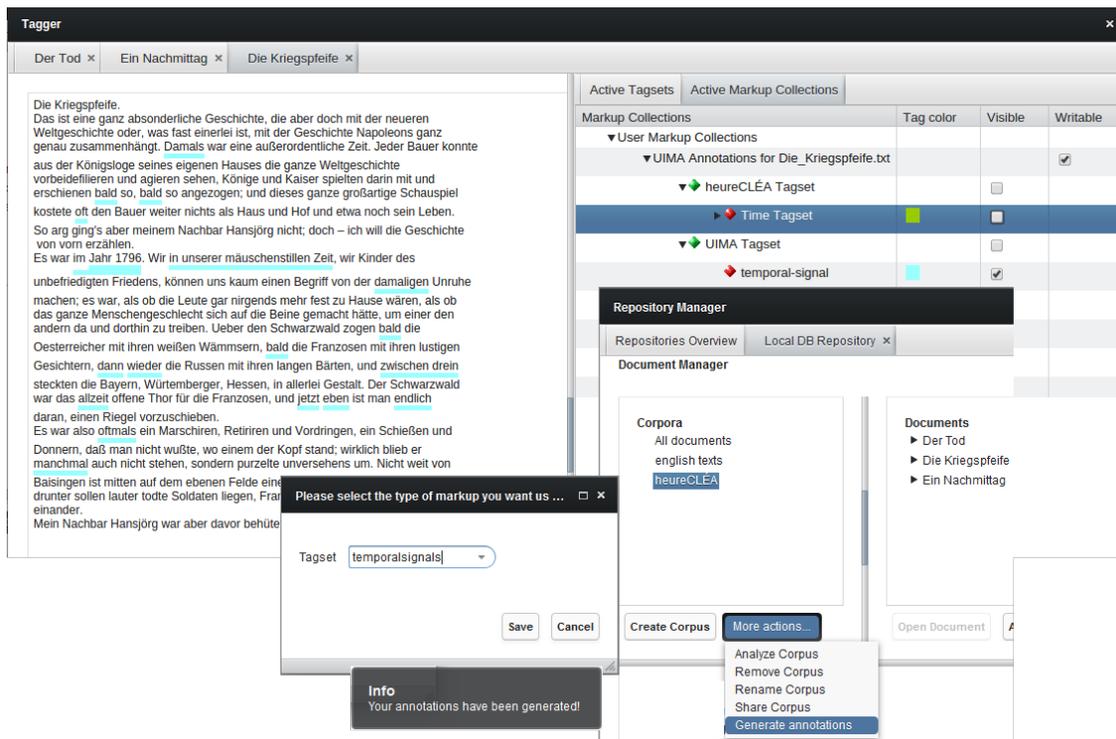


Figure 2: Screenshot showing automatic annotations within CATMA.

Abstract zum Vorhaben „Sprachwissenschaftliche Untersuchungen zum Klagspiegel Conrad Heydens (1436) und zum Laienspiegel Ulrich Tenglers (1511)“

Von Dr. Barbara Aehnlich und Elisabeth Witzenhausen

Das Forschungsvorhaben ist interdisziplinär angelegt und beruht auf einem Korpus von verschiedenen Textzeugen zweier frühneuhochdeutscher Rechtsbücher des 15. und 16. Jahrhunderts, Klagspiegel und Laienspiegel. Der Klagspiegel ist das mit Abstand älteste populärwissenschaftliche Rechtsbuch der Rezeptionszeit und bildet mit dem Laienspiegel zusammen die wichtigste Grundlage an rechtswissenschaftlichen populären Texten des 15. und 16. Jahrhunderts. Ziel ist die Untersuchung der sprachlichen Besonderheiten der Texte und ihrer Auswirkungen auf die Rezeptionsgeschichte des römischen Rechts in Deutschland. Neben der korpusbasierten linguistischen Analyse der Bücher, die eine völlig neue Textsorte begründen, bietet das Projekt auch aus der Perspektive der historischen Rechtssprachenforschung einen innovativen Ansatz. Das Erkenntnisinteresse liegt hierbei auf der Geschichte von Kulturtransferprozessen innerhalb der Jurisprudenz. Durch semantische und linguistische Annotationen wird eine umfassende Forschungsgrundlage geschaffen, die für die Schließung rechts- und sprachhistorischer Forschungslücken einen zentralen Beitrag leistet. Ein weiterer Schritt soll die Digitalisierung mehrerer Ausgaben des Klagspiegels sein, um Prozesse des Schreibsprachwandels im 15. und frühen 16. Jahrhundert nachzuvollziehen. Bisher gibt es kein Korpus frühneuhochdeutscher Rechtstexte.

In einem ersten Schritt zur Vorbereitung des Projektes wurden verschiedene Annotationstools getestet und geeignete Formate für die Speicherung evaluiert. Aktuell werden mit der Jenaer Computerlinguistik Möglichkeiten der Normalisierung und automatischen Annotation erprobt. Ziel ist die Beantragung eines größeren Forschungsprojektes, das bestehende Werkzeuge nutzt und die Technologie auf die Besonderheiten des Rechtskorpus anpasst. Das Poster soll die bisherigen methodologischen Überlegungen und Probleme darstellen und bietet somit gleichzeitig einen Überblick und eine Evaluation der aktuell zur Verfügung stehenden Open Source Software zu Annotationszwecken.

Die Untersuchungen beziehen sich zum einen auf die sprachliche Herkunft des Klagspiegels und des Laienspiegels. Es soll festgestellt werden, welche Textsorte mit welchen spezifischen

sprachlichen Eigenheiten vorliegt. Zudem muss geklärt werden, ob diese Rechtsbücher aufgrund ihrer Herkunft nur im südwestdeutschen Raum oder aber im gesamten hochdeutschen Sprachgebiet verständlich waren. Dabei wird nach möglichen Ausgleichstendenzen gesucht, die vom Oberdeutschen abweichen. Auf der Ebene der Syntax ist zu fragen, welche Strukturen die sprachliche Einfachheit und leichte Verständlichkeit ausmachen, die den Texten in der gesamten (bisher ausschließlich juristischen) Forschung zugeschrieben wird. Im Bereich des Wortschatzes sind besonders die Bezeichnungen juristischer Fachbegriffe oder Tatbestände für die Forschung interessant, denn für diese gab es zuvor im Deutschen keine entsprechenden Termini. Zum anderen soll untersucht werden, inwieweit Klagspiegel und Laienspiegel frühneuhochdeutschen Sprachstandard aufweisen und ob die beiden Bücher durch ihre Verbreitung eine wesentliche Rolle für die Entwicklung des neuhochdeutschen Sprachstandards im Rahmen rechtswissenschaftlicher Prozesse gespielt haben können. Ein Vergleich mehrerer Textzeugen der Rechtsbücher liefert Erkenntnisse des frühneuhochdeutschen Schreibsprachwandels. Der Einfluss der beiden Texte auf die deutsche Standardsprache sowie auf die deutsche Rechtssprache wurde bisher noch nicht analysiert; das Vorhaben soll hierfür eine nutzbare Ausgangsbasis liefern. Eine zentrale Frage ist dabei auch, inwieweit römisches Recht und deutsches Recht sprachlich unterschiedlich vermittelt wurden und ob textinternen Varianz zwischen den einzelnen Passagen, die zum Teil auch literarisiert sind, festzustellen ist.

Zwei Textzeugen, jeweils eine Ausgabe des Laien- und des Klagspiegels, liegen bereits in digitalisierten Abbildungen vor und wurden transkribiert. Im nächsten Schritt werden sie in ein XML-Format übertragen und sollen semantisch sowie linguistisch annotiert werden, um eine valide Datenbasis für die Untersuchung zu schaffen und das Korpus in einem standardisierten Format in einer Infrastruktur der Digital Humanities zur Verfügung stellen zu können. Im Sinne eines vielseitig nutzbaren Korpus soll die Transkription diplomatisch, mit allen Sonderzeichen und typografischen Besonderheiten, abgebildet werden. Problematisch ist die heterogene Gestalt der Texte, die Mehrfachannotationen notwendig macht. Alle Annotationen werden deshalb in einem XML-stand-off Format vorgenommen, um eine leichte Übertragung in andere Formate und einen annotationsfreien Primärtext zu ermöglichen. Das TCF-Format bietet hierfür eine gute Möglichkeit und ist mit vielen anderen Formaten kompatibel.¹ Werkzeuge wie WebAnno² oder GATE³ bieten geeignete Arbeitsoberflächen, deren Vor- und Nachteile es zu diskutieren gilt.

¹ http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format. (06.10.2014, 13.30 Uhr).

² <https://code.google.com/p/webanno/>. (06.10.2014, 13.46 Uhr).

Die Präsentation stellt somit zum einen den Mehrwert der bisher geleisteten Forschungsarbeit im Rahmen der Digital Humanities für sprachwissenschaftliche Untersuchungen historischer Texte heraus, zum anderen werden Grenzen in der Annotation heterogener und nicht standardisierter Sprachdaten deutlich, die weiterer Forschungsarbeit bedürfen. Die interdisziplinär angelegte Forschungsfrage und die unterschiedlichen Zielgruppen des zu erstellenden Korpus sind Faktoren, die es bei der Aufarbeitung der Daten besonders zu beachten gilt.

³ <https://gate.ac.uk/sale/tao/split.html>. (06.10.2014, 12.51 Uhr).

Poster-Abstract

Erfahrungen aus dem *Bibliotheca legum*-Projekt. Zum Aufbau einer Handschriftendatenbank

(<http://www.leges.uni-koeln.de>)

Daniela Schulz und Dominik Trump, M.A. (Universität zu Köln)

Seit 2012 entsteht am Lehrstuhl für die Geschichte des Mittelalters (Prof. Dr. Karl Ubl) in Köln eine Datenbank, welche die handschriftliche Überlieferung des weltlichen Rechts im Früh- und Hochmittelalter in den Blick nimmt. Unter weltlichem Recht werden dabei sowohl das römische Recht als auch die germanischen Volksrechte, die sog. *leges barbarorum*, verstanden. Durch die genaue Erfassung aller relevanten Textzeugen, ihrer Produktion und Verbreitung ist es möglich, Rückschlüsse auf das damalige Rechtswissen zu ziehen.

Das Projekt versteht sich zum einen als Ergänzung zur *Bibliotheca capitularium* von Hubert Mordek¹, der grundlegend alle Handschriften mit fränkischen Herrschererlassen, den sog. Kapitularien, gesammelt hat.² Diese bilden eine weitere zentrale Quelle der frühmittelalterlichen Rechtsgeschichte. Zum anderen bietet die *Bibliotheca legum* einen umfassenden Überblick über die Überlieferung und damit Rezeption des römischen Rechts im frühen Mittelalter – ein Gebiet, welches bisher in der Forschung nur relativ wenig Beachtung gefunden hat.

Die *Bibliotheca legum* bietet momentan zu 296 Handschriften Informationen zu Datierung, Entstehungsort, Provenienz, äußerer Beschreibung, Inhalt, Literatur und vor allem zu im Internet frei zugänglichen Ressourcen wie Digitalisaten (z.B. aus *Europeana regia*, *Gallica* oder der Bayerischen Staatsbibliothek München) und Katalogeinträgen (z.B. *Manuscripta Mediaevalia*). Das Projekt ist somit konzeptionell als Portal angelegt, welches nicht nur selbst umfassende Informationen bietet, sondern auch vorhandene Ressourcen nachnutzt und miteinander verknüpfen möchte.

Für das Projekt wurden sowohl die einschlägigen Editionen der Rechtstexte als auch ältere und insbesondere neuere und neueste Forschungsliteratur systematisch ausgewertet. Die gesammelten Informationen, die zunächst nur für eine lehrstuhlinterne Nutzung vorgesehen waren, wurden dabei anfänglich in einer Word-Tabelle gesammelt. Um die Ergebnisse in Form einer Webpräsenz einem breiteren Publikum zugänglich machen zu können, überführte

¹ Hubert Mordek, *Bibliotheca capitularium regum Francorum manuscripta. Überlieferung und Traditionszusammenhang der fränkischen Herrschererlasse* (MGH Hilfsmittel 15), München 1995.

² Aufgrund der freundlichen Genehmigung der *Monumenta Germaniae Historica* (MGH) in München, kann seine Studie bzw. Auszüge aus dieser (auf den einzelnen Handschriftenseiten) zum Download angeboten werden.

man diese Datensammlung nach XML und generierte daraus Handschriftenbeschreibungen nach TEI P5, welche den aktuellen Forschungsstand abbilden und zum Download verfügbar sind.

Zur Verwaltung der Webpräsenz wurde mit Wordpress ein kostenfreies Content Management System gewählt, welches zwar als Blogsoftware weite Verbreitung im World Wide Web gefunden hat, bisher aber nicht für XML-basierte Digital Humanities-Projekte herangezogen wurde. Gerade wegen der breiten Community, die an diesem CMS partizipiert und damit dem Vorhandensein zahlreicher Plugins zur Erweiterung der Funktionalitäten, hat sich Wordpress für das Projekt, welches weder über Drittmittel noch über einen technischen Partner verfügt, insgesamt als sehr geeignet erwiesen. Aufgrund der positiven Erfahrungen folgt die momentan entstehende Webpräsenz der Arbeitsstelle „Edition der fränkischen Herrschererlasse“ (ein Projekt der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste unter der Leitung von Prof. Dr. Karl Ubl; URL: <http://capitularia.uni-koeln.de/>) dem Vorbild der *Bibliotheca legum* und setzt ebenfalls auf Wordpress auf.

Die *Bibliotheca legum* bietet die folgenden *Features*:

- Mehrsprachigkeit (Menüführung und Inhalt in deutscher und englischer Sprache)
- verschiedene Browsingzugänge (z.B. nach Signaturen, enthaltenen Rechtstexten, Entstehungszeit und -ort) zu den fast 300 Handschriftenbeschreibungen
- Volltextsuche und facetiierte Suche
- Inter- und Hyperlinking (externe Ressourcen)
- Einleitungstexte
- Übersicht über die in mittelalterlichen Bibliothekskatalogen bezeugten Rechtscodices
- umfangreiche Projektbibliographie sowie Register zu Orten, Personen und haltenden Institutionen unter Verwendung von Normdaten wie VIAF und TGN
- ein Projektblog (deutsch/englisch), der über aktuelle Entwicklung informiert
- umfangreiche Materialsammlung und Visualisierungen (Karten, Stemmata, Transkriptionen)
- Downloads (z.B. die *Bibliotheca capitularium*, Handschriftenbeschreibungen)

Innerhalb recht kurzer Zeit – die Datenbank ist erst seit September 2012 online – konnte sie sich in der historischen und rechtshistorischen Forschung als Arbeitsinstrument etablieren. Neben dem regelmäßigen Einsatz in Lehrveranstaltungen an der Universität zu Köln, fand sie z.B. auch im MOOC „Karl der Große – Pater Europae?“ (URL:

<https://iversity.org/de/courses/karl-der-grosse-pater-europae>) von apl. Prof. Dr. Rainer Leng (Universität Würzburg) Erwähnung, wo sie als Arbeitsinstrument für die Geschichtswissenschaft präsentiert wird.

Das Poster soll das Projekt nun erstmals auch der deutschsprachigen DH-Community vorstellen. Dabei soll zum einen der aktuelle Stand des Projekts sowie dessen Nutzen für die Wissenschaft, zum anderen dessen – sicher nicht gewöhnliche – Genese dargestellt werden. Die Präsentation kann für all jene von Interesse sein, die digitale Projekte unter ähnlichen Umständen bzw. Voraussetzungen (Fehlen technischer und finanzieller Mittel) realisieren wollen.

Leitung des Projekts: Prof. Dr. Karl Ubl, Lehrstuhl für die Geschichte des Mittelalters, Schwerpunkt Früh- und Hochmittelalter, Historisches Institut der Universität zu Köln

Mitarbeiter: Daniela Schulz (Technische Umsetzung), Dominik Trump, M.A. (Inhaltliche Bearbeitung)

Common Names 4 living organisms @ EUROPEANA

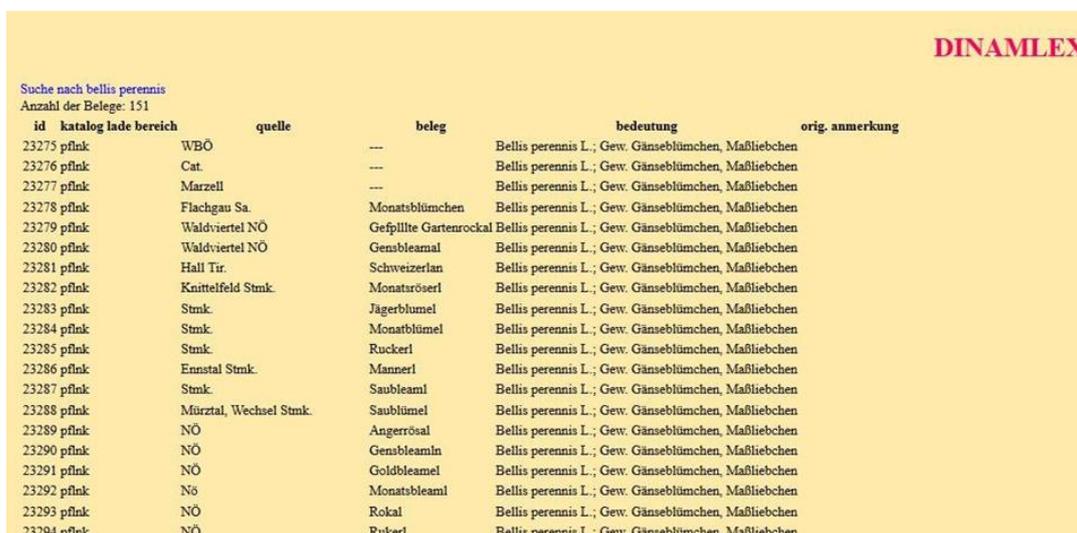
Ein Beispiel für Mehrwert durch interdisziplinäre Kollaboration

Im Zeitalter von digitalisierten Wissenschaften, in denen das Aufbereiten, das zur Verfügung stellen und Austauschen von Daten, Fakten und Erkenntnissen einen mehr als wichtigen Stellenwert einnimmt, ist es somit nicht verwunderlich, dass auch der Aspekt der interdisziplinären Zusammenarbeit und die Gestaltung bzw. Durchführung von Landesgrenzen übergreifenden Projekten mehr und mehr in das Zentrum der Arbeit von Forschern und Forscherinnen rückt und gleichzeitig diese Anforderungen an die Wissenschaften selbst gestellt werden.

Im vorgeschlagenen Poster wird eine interdisziplinäre Zusammenarbeit zwischen den Lexikographen und Lexikographinnen des *Instituts für Corpus Linguistik und Texttechnologie* der *Österreichischen Akademie der Wissenschaften* mit den Kollegen des *Common Names Service* (CNS) des Naturhistorischen Museums Wien (NHM) vorgestellt. Der Mehrwert für beide Disziplinen, die Datenpublikation im Rahmen von Europeana.EU und die Europäisierung der Services stehen im Zentrum der Präsentation.

Die Zusammenarbeit war seitens der Projektpartner wie folgt motiviert:

- 1) Die *Datenbank der bairischen Mundarten in Österreich* (DBÖ) weist unter anderem eine umfassende Sammlung volkstümlicher Pflanzennamen auf (geschätzt 30,000). Um die Daten lexikographisch im *Wörterbuch der bairischen Mundarten in Österreich* (WBÖ) entsprechend wissenschaftlich dokumentieren zu können, muss der Lexikograph bzw. die Lexikographin den jeweiligen Common Name einer konkreten Pflanze zuweisen (Definition). Aufgrund der Historizität der Daten (Sammelzeitraum „von den Anfängen bis in die Gegenwart“) ist das eine fachspezifische Aufgabe der Botaniker und Botanikerinnen.



Suche nach bellis perennis
Anzahl der Belege: 151

id	katalog lade bereich	quelle	beleg	bedeutung	orig. anmerkung
23275	pflnk	WBÖ	---	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23276	pflnk	Cat.	---	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23277	pflnk	Marzell	---	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23278	pflnk	Flachgau Sa.	Monatsblümchen	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23279	pflnk	Waldviertel NÖ	Gefüllte Gartenrockal	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23280	pflnk	Waldviertel NÖ	Gensbleamal	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23281	pflnk	Hall Tir.	Schweizerlan	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23282	pflnk	Knittelfeld Stmk.	Monatsröserl	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23283	pflnk	Stmk.	Jägerblumel	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23284	pflnk	Stmk.	Monatsblümel	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23285	pflnk	Stmk.	Ruckerl	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23286	pflnk	Ennstal Stmk.	Mannerl	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23287	pflnk	Stmk.	Saubleam	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23288	pflnk	Mürztal, Wechsel Stmk.	Saublümel	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23289	pflnk	NÖ	Angerrösal	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23290	pflnk	NÖ	Gensbleamln	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23291	pflnk	NÖ	Goldbleamel	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23292	pflnk	Nö	Monatsbleam	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23293	pflnk	NÖ	Rokal	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	
23294	pflnk	NÖ	Ruckerl	Bellis perennis L.; Gew. Gänseblümchen, Maßliebchen	

- 2) Umgekehrt ist es für die Bestimmung einer Pflanze und deren Zuweisung im Zeit-Raum-Kontinuum sehr hilfreich für die Botaniker und Botanikerinnen, auf eine umfassende Sammlung wie die DBÖ zurückgreifen zu können: Daraus können Varianten für Volkssprache, Büchernamen und historische Taxonomien ebenso gewonnen werden wie Rückschlüsse auf Verwendung (Ethnobotanik) und auf Verbreitung (Georeferenzierung) getätigt werden.

DATENBANK DER
BAIRISCHEN MUNDARTEN
IN ÖSTERREICH @ ELECTRONICALLY MAPPED

Abfrage Projekt Ressourcen

Index

Index

- Lemma
- Bibliographie
- Personen
- Orte
- Gemeinden
- Regionen
- Fragebuch

Beleg

Rehling

Details

- **Hauptlemma:** [Rehling](#)
- **Wortart:** Substantiv
- **Kategorie:** Name
- **Belegtyp:** Lautung

Erklärung

- **Bedeutungen:** Eierschwamm oder Gelbling, der Eierpilz

Belegstelle

- **Belegt für:** Gmunden
- **Belegt laut:** [Weinkopf \(1926\)](#)

Darstellung in der DBÖ

Um die Motivationen der beiden kollaborierenden Einheiten herzustellen, wurden folgende Schritte eingeleitet:

- 1) Entwicklung eines allgemeinen Schemas zur Modellierung von Pflanzennamen und zugehöriger Informationen.
- 2) Modellierung der Pflanzennamen in SKOS.
- 3) Datenaufbereitung in der DBÖ.

Im Zuge der Datenaufbereitung übernimmt eine überregionale bzw. standardnahe Bezeichnung die Funktion des Hauptlemmas, unter welchem die jeweiligen regionalsprachlichen bzw. dialektalen Entsprechungen in Kombination mit den biologischen Informationen der botanischen Datenbank zusammengefasst werden.

Die dabei entstehenden Datensätze stellen eine sehr umfangreiche, wissenschaftlich fundierte Sammlung sprachlicher sowie sozio-kultureller Phänomene dar, womit sie als Quelle für verschiedenste wissenschaftliche Forschungsfragen herangezogen werden können. Diese bilden demnach die Grundlage für eine institutsübergreifende Zusammenarbeit und Verkettung von Daten, sowie deren Bereitstellung für die öffentliche Verwendung.

- 4) Entwicklung eines Webservices zur Kommunikation zwischen den existierenden Datenbanken zwecks abfragegesteuertem Datenaustauschs.
- 5) Publikation der Daten in Europeana im Kontext des Projekts OpenUp!

The screenshot shows the Europeana interface for the entry 'Cantharellus cibarius'. It features a drawing of the mushroom, a list of synonyms in various languages, and metadata including the identifier 'ETI - IGMF - 360', the relation to the Biodiversity Library, and the source 'Mushrooms and other Fungi'. The provider is listed as 'OpenUp!' and the providing country is 'Netherlands'.

Darstellung in Europeana

Das Projekt BioLing wird im Kontext der COST Aktion IS1305 weitergeführt:

Durch Berücksichtigung weiterer lexikographischer Ressourcen im Webservice (und somit letztlich in Europeana) sollen Möglichkeiten geschaffen werden, europäisches Kulturgut besser zugänglich zu machen und Zusammenhänge in der Benennungsmotivik zu erarbeiten. Im Zuge dessen wird die interdisziplinäre Kommunikation zwischen den einzelnen wissenschaftlichen Instituten innerhalb Österreichs intensiviert und es wird dazu beitragen, dass es auf internationaler Ebene vermehrt zu Kooperationen kommt und länder-, instituts- und wissenschaftsübergreifende Projekte konzipiert werden.

Als Ausblick basierend auf der kollaborativen Zusammenarbeit und der bestehenden Infrastruktur: Über die Aufarbeitung der jeweiligen Etymologie einer

Pflanzennamenbezeichnung werden tieferliegende Zusammenhänge erarbeitet, die als Konzepte im Europäischen Kontext qualifizierbar und quantifizierbar gemacht werden sollen.

Im Kontext von DARIAH-EU wird auf Basis der etablierten modellhaften Zusammenarbeit zwischen ÖAW und NHM ein Beitrag traditionell lexikographischer europäischer Arbeiten zu Controlled Vocabularies erarbeitet.

Digital Humanities in der Hochschullehre – Erfahrungen aus dem Lern-Lehr-Projekt **„Digitale Medien in den Geisteswissenschaften in Lehre und Forschung“**

Patrick Pfeil, M.A. (Alte Geschichte, Universität Leipzig), Sabrina Herbst, M.A. (Medienzentrum, TU Dresden), Corina Willkommen, B.A. (Alte Geschichte, Universität Leipzig)

Die zunehmende Bedeutung der Digital Humanities für die Geisteswissenschaften erfordert auch die Vermittlung neuer Kompetenzen an Studierende und damit die Konzeption neuartiger Lehr- und Lernangebote. In diesem Zusammenhang ist das vom BMBF geförderte und vom Projektverbund „Lehrpraxis im Transfer“ betreute Lern-Lehr-Projekt zu sehen (<https://www.hds.uni-leipzig.de/index.php?id=projektkohorte-3>). Am Vorhaben beteiligt sind die Alte Geschichte der Universität Leipzig (Prof. Charlotte Schubert), die Korpuslinguistik der TU Dresden (Prof. Joachim Scharloth) und das Medienzentrum der TU Dresden (Prof. Thomas Köhler). Ziel des Projekts ist die Einbindung von Lehrangeboten der Digital Humanities (konkret hier die Projekte eAQUA, eComparatio und Papyrusportal Deutschland) in die Regellehre des Bachelor Studienganges Geschichte und des Master Studienganges Klassische Antike der Universität Leipzig sowie die Entwicklung von Selbstlernmodulen zur Einführung in die Korpuslinguistik an der TU Dresden.

Im Vortrag wird zunächst das Vorhaben im Einzelnen vorgestellt. Im Anschluss werden die gemachten Erfahrungen mit der Einbindung in die Regellehre aus Sicht der Dozierenden dargestellt. Hierbei wird schwerpunktmäßig mit den im Wintersemester 2014/15 abgehaltenen zwei Digital-Humanities-Modulen an der Universität Leipzig gearbeitet. Auf das Selbstlernmodul der TU Dresden wird schlaglichtartig eingegangen. Zum Abschluss des Vortrages steht die Sicht der Studierenden im Mittelpunkt der Ausführungen.

Durch die Integration der Digital Humanities in die Regellehre der verschiedenen Geisteswissenschaften ändert sich das Profil der Studiengänge. Dies bringt neue Anforderungen an Lehrende und Studierende mit sich. Es gilt zu fragen, wie heutige Studierende, die Angebote aus den Digital Humanities annehmen und für das eigene Studium nutzbar machen. Stellen diese dabei ein Zusatzangebot dar oder gehen die Möglichkeiten der Digital Humanities in das alltägliche Arbeitsgerüst der Studierenden ein, wie es früher beim Wörterbuch oder bei einer Grammatik der Fall war? Wie entwickelt man bei den Studierenden die Bereitschaft sich auf Angebote aus den Digital Humanities einzulassen und welche Methoden sind dabei anwendbar? Darüber hinaus ist von Interesse, in welcher Phase des Studiums man diese Angebote einbringen sollte und ob man durch Digital Humanities die Befähigung der Studierenden zum Forschenden Lernen besonders fördern kann.

In den letzten Jahren konnten an der Universität Leipzig mehrere Seminare im Bereich Digital Classics angeboten werden, die aufeinander aufbauen und seit vier Jahren durch dieselben DozentInnen veranstaltet werden. Die Erfahrungen der DozentInnen bezüglich der didaktischen Umsetzung konnten durch mehrere Förderprojekte verstetigt werden, mit dem Ziel, die sich entwickelnden digitalen Methoden dauerhaft in den Hochschulunterricht einzupflegen. Im Rahmen besagter Projekte konnten außerdem parallel zu den Lehrveranstaltungen weiterbildende Maßnahmen angeboten werden, wie Workshops zu Programmierung, Visualisierung und Digitaler Edition. Aufbauend auf diesen Erfahrungen stehen das derzeit durchgeführte Bachelorseminar zur „Einführung in die antike Numismatik“ und das Masterseminar „Zur kulturellen Praxis des Zitierens“ in der Tradition der Digital Classics-Seminare und profitiert nicht nur aus den Erfahrungen der DozentInnen, sondern auch durch die Studienarbeiten vorangegangener Seminare, die in einer eigenen

Publikationsreihe „eAQUA Working Papers“ erschienen sind (<http://journals.ub.uni-heidelberg.de/index.php/eaqua-wp>). Aufgrund dieser Arbeitsgrundlage haben die Studierenden die Möglichkeit auf die in den vorangegangenen Veranstaltungen entwickelten Fragestellungen, Lösungsstrategien, Fehleranalysen und methodischen Entwicklungen zurückzugreifen und sich neu zu orientieren.

Ziel des Projektes ist zum einen der selbstständige und praktische Umgang mit digitalen Tools, um alternativ und ergänzend zu den klassischen Fachmethoden Lösungsstrategien zur Bearbeitung historischer Fragestellungen zu entwickeln. Zum anderen steht das Konzept des Forschendes Lernens im Fokus der Seminare.

Die Einführung in die digitalen Tools des Programms eAQUA („Extraktion von strukturiertem Wissen aus antiken Quellen“ - <http://www.eaqua.net/>) „Kookkurrenzanalyse“, „Zitationsgraph“ und „Mental Maps“ erfolgt dabei durch die DozentInnen anhand fachspezifischer Fragestellungen. In einem weiteren Schritt erlernen Studierende die praktische Handhabung der Tools mittels eigens dafür erstellter Übungshandbücher im iBook-Format. Vor allem der praktische Umgang mit den vorgestellten Tools unter Einbeziehung eigener Fragestellungen hat sich als überaus erfolgreich erwiesen, als es darum ging, die Studierenden zur aktiven Mitarbeit anzuregen. In diesem Fall konnten die Studierenden als „ExpertInnen“ eines bereits behandelten Themas und mit dem Wissen um das Ergebnis ihrer Fragestellung, selbige durch das Tool verifizieren lassen. Die Erwartungshaltung der Studierenden konnte durch Einsatz digitaler Tools bestätigt und erheblich ergänzt werden. Die Vorteile des Einsatzes von digitalen Hilfsmitteln zeigten sich besonders deutlich im Vergleich verschiedener Arbeitsmethoden, die zur Bearbeitung wissenschaftlicher Fragestellungen herangezogen worden. Die Studierenden entwickeln in einer teils autodidaktischen Atmosphäre nicht nur fachliche Kompetenzen, sondern konnten auch gruppendynamisch in einen Diskurs treten und damit soziale Kompetenzen erwerben. Die Ergebnisanalyse wird direkt im Unterricht von den KommilitonInnen in erster und nachfolgend von den DozentInnen in zweiter Instanz vorgenommen.

Im zweiten Teil des Vortrags wird die Perspektive der Studierenden betrachtet. Dabei gilt es verschiedenste Herausforderungen zu überwinden: Dies ist zum einen die vielfach diskutierte Diskrepanz zwischen dem Nutzungsverhalten Neuer Medien der Generation der sog. „Digital Natives“ (Prensky 2001; 2001a - zur kritischen Auseinandersetzung mit dem Begriff der Digital Natives u. a.: Arnold & Weber 2013) und ihrem Einsatz Neuer Medien für das Studium (vgl. hierzu Weller et al. 2014). Zum anderen müssen Lehrangebote konzipiert werden vor dem Hintergrund einer hohen Diversität der Zielgruppe, hinsichtlich fachlicher Hintergründe und Motivation der Studierenden (Schwerpunktmodul oder Wahlpflichtbereich), Studiengang (BA oder MA) und Fachsemester. Es ist daher in unterschiedlicher Hinsicht von Bedeutung, bei der Konzeption von Lehrangeboten unter Einbezug digitaler Technologien die unterschiedlichen Bedürfnisse der Studierenden in den Blick zu nehmen. Nicht zuletzt handelt es sich bei der Einführung von Digital Humanities-Lehrangeboten um eine Lerninnovation (Kerres 2013) in der geisteswissenschaftlichen Lehre, bei deren Einführung besondere Akzeptanz durch die studentische Zielgruppe notwendig ist, um Lernerfolge zu erzielen und eine Verstetigung zu ermöglichen. Bei der Planung von Lehrangeboten bietet sich daher im Vorfeld die Durchführung einer Anforderungsanalyse an, um möglichst viel über die Zielgruppe der Lernenden und die das Lernen beeinflussenden Rahmenbedingungen herauszufinden. Dabei geht es bei der Konzeption eines Digital Humanities-Lehrangebotes vor allem um unterschiedliche Nutzungsgewohnheiten, Einstellungen und Kompetenzen hinsichtlich Neuer Medien bei den Studierenden, ihre Erfahrungen mit unterschiedlichen Lehr- und Lernformaten sowie strukturelle Einflussfaktoren, wie Studiengang, Modulform, Fachsemester zu identifizieren. Für das im Lehr-Lern-Projekt „Neue Medien in den Geisteswissenschaften in Lehre und Forschung“ zu entwickelnde Lehrangebot im Bereich der

Digital Humanities wurden im Mai 2014 gemeinsam mit den Studierenden Anforderungen und Praktiken des Einsatzes Neuer Medien in den Geisteswissenschaften erhoben. Eine ebenso wichtige Rolle wie die Durchführung einer Anforderungsanalyse spielt außerdem die Evaluierung des Lehrangebotes im Nachgang, so dass das im Wintersemester 2014/2015 erprobte Lehrangebot an der Universität Leipzig daher auch Ende 2014/ Anfang 2015 evaluiert wird.

Sowohl für die Anforderungsanalyse im Vorfeld als auch für die Evaluation wurde die Fokusgruppe als Erhebungsinstrument der empirischen Sozialwissenschaft gewählt. Fokusgruppeninterviews eignen sich aufgrund der breiten kollektiven Wissensbasis der TeilnehmerInnen besonders, um unterschiedliche Facetten einer Problemstellung zu erheben (vgl. Schulz 2010).

Diese erste Fokusgruppe (Anforderungsanalyse) wurde mit 11 Studierenden des Seminars „Digitale Altertumswissenschaft“ im vergangenen Sommersemester an der Universität Leipzig durchgeführt. Hierfür wurden die Studierenden einerseits zu den generellen Rahmenbedingungen ihres Lernens im Studium, andererseits zu ihren Erfahrungen mit der Seminarstruktur des Seminars „Digitale Altertumswissenschaft“ sowie den dort vorgestellten Digital Humanities-Werkzeugen „Perseus-Datenbank“ (www.perseus.tufts.edu) und den eAqua-Tools „Kookkurenzanalyse“, „Zitationsgraph“ und „Mental Maps“ befragt. Die Abschlussphase des Interviews diente dazu, die Bereitschaft der Studierenden, sich weitere Kompetenzen im Bereich der Digital Humanities anzueignen, auszuloten und Verbesserungsvorschläge für die Vermittlung von Inhalten in der Lehre zu erhalten. Die so erhobenen Anforderungen wurden dann für die Entwicklung des Lehrangebotes genutzt. Dabei hat sich unter anderem gezeigt, dass die Studierenden zwar an dem Erwerb von Kompetenzen im Bereich der Digital Humanities interessiert sind, jedoch über ein nur sehr rudimentäres Verständnis des Begriffs der Digital Humanities verfügen. Es wurde ebenfalls deutlich wie wichtig eine Einführung in die verschiedenen Methoden der Digital Humanities ist und welche Rolle bestimmte Lehrformate dabei spielen. Die Durchführung des zweiten Fokusgruppeninterviews zur Evaluation des im Wintersemester 2014/2015 durchgeführten Lehrangebots an der Universität Leipzig ist für Ende 2014/ Anfang 2015 geplant, um zu erheben, wie gut sich dieses in der Praxis bewährt hat.

Im Vortrag werden die Ergebnisse beider Fokusgruppeninterviews vorgestellt und die daraus abgeleiteten Handlungsempfehlungen für die Erarbeitung des Lehrangebots präsentiert. Darüber hinaus sollen die Ergebnisse der Usability-Untersuchung des Lehrangebots Ende 2014 dargestellt und die Frage diskutiert werden, inwieweit eine Berücksichtigung der Anforderungen der Studierenden gelungen ist.

Literatur:

Arnold, P. & Weber, U. (2013): Die „Netzgeneration“. Empirische Untersuchungen zur Mediennutzung bei Jugendlichen. In: M. Ebner & S. Schön (Hrsg.), L3T. Lehrbuch für Lernen und Lehren mit Technologien. Online-Dokument: <http://l3t.eu/homepage/das-buch/ebook-2013/kapitel/o/id/144/name/die-netzgeneration> (06.11.2014)

Kerres, M. (2013): Mediendidaktik, Konzeption und Entwicklung mediengestützter Lernangebote. München, Oldenbourg.

Prensky, M. (2001): Digital Natives, Digital Immigrants. In: On the Horizon 9,5 (2001). Online-Dokument: <http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf> (10.11.2014)

Prensky, M. (2001a) Digital Natives, Digital Immigrants. Part II. Do They Really Think Differently? In: On the Horizon 9,6 (2001). Online-Dokument: <http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part2.pdf> (10.11.2014)

Schulz, M. (2012): Quick and easy?! Fokusgruppen in der angewandten Sozialwissenschaft. In: M. Schulz, B. Mack & O. Renn, Fokusgruppen in der empirischen Sozialwissenschaft. Von der Konzeption bis zur Auswertung. Wiesbaden, S. 9-22.

Ansprechpartner:

Patrick Pfeil, M.A.

Universität Leipzig, Fakultät für Geschichte, Kunst und Orientwissenschaften, Historisches Seminar, Lehrstuhl für Alte Geschichte, Projektleiter Lern-Lehr-Projekt „Digitale Medien in den Geisteswissenschaften in Lehre und Forschung“

GWZ, Beethovenstraße 15, 04107 Leipzig (Raum: H4 2.16)

Tel.: +49 341 9737077, Fax: +49 341 9737071

Email: ppfeil@uni-leipzig.de

Posterpräsentation

Jahrestagung DHd 2015

eComparatio

Editionsvergleich

Oliver Bräckel, Hannes Kahl, Friedrich Meins, Charlotte Schubert

Das von der Deutschen Forschungsgemeinschaft (DFG) geförderte Projekt eComparatio wird seit 2014 als Kooperationsprojekt des Lehrstuhls für Alte Geschichte der Universität Leipzig und des ICE (Interdisciplinary Center of E-Humanities in History and Social Sciences/ Forschungsstelle am Max-Weber-Kolleg für kultur- und sozialwissenschaftliche Studien an der Universität Erfurt) entwickelt. Das Ziel des Projektes ist es, eine modular aufgebaute Anwendung zu entwickeln, die es ermöglicht, verschiedene Versionen eines Textes (aus Handschriften, gedruckten oder digitalen Texteditionen) miteinander zu vergleichen. Das Kernstück der Anwendung ist ein Modul zum Vergleich von Textausgaben, das auch die Erstellung eines Variantenapparates für digitale Editionen antiker Autoren ermöglicht. Die Zahl der Vergleichstexte ist beliebig, ebenso das Eingabeformat (TXT, HTML, XML, JSON, PDF). Die Anwendung wird frei skalierbar sein, so dass der Umfang der zu vergleichenden Texte nicht beschränkt ist, das Ergebnis (Kollationierung) soll in Form von Listen als kritischer Apparat (positiver oder negativer Apparat) oder auch in beliebiger anderer Form ausgegeben werden können. In einem weiteren Modul soll für Autorenreferenzen bei der Abfrage von online-Datenbanken die Anbindung an das Referenzsystem CTS (Canonical Text Services) und die Referenz auf Images von Handschriften (über das Image Citation Tool der CITE Collection Services) ermöglicht werden. Die Ansprechbarkeit für weitere Adressschemata wird ebenfalls implementiert (z.B. für JSON und den im Aufbau befindlichen PID-Service von CLARIN-D). Im bisherigen Verlauf des Projektes ist es gelungen, die Grundfunktionen des Tools zu implementieren und es in die Lage zu versetzen eine beliebig große Anzahl an Texten miteinander zu vergleichen. Dabei sind drei unterschiedliche Ansichten entstanden, die es dem Benutzer ermöglichen das Ergebnis aus verschiedenen Perspektiven zu betrachten. Die Detailansicht zeigt einen Text und markiert entsprechende Unterschiede zu anderen Texten. Die Parallelansicht (siehe auch Abbildung) zeigt alle Texte nebeneinander und markiert die Unterschiede farblich. Die Buchansicht schließlich zeigt wieder nur einen Text an und visualisiert

die Varianten im Stile traditioneller Printeditionen unter dem betreffenden Abschnitt. Zu betonen ist dabei, dass der Ausgangstext für den Vergleich bei jeder Ansicht frei wählbar ist und sich somit nicht auf einen zu bevorzugenden Haupttext festgelegt bzw. eine Gewichtung der Textzeugen vorgenommen wird.

Die Visualisierung und Ergebnissicherung ermöglicht zum einen, einen schnellen Überblick über die Text- und Editions-geschichte verschiedener in digitalisierter Form vorliegender Werke zu erlangen. Darüber hinaus eignet sich das Tool als Hilfsmittel zum Kollationieren bei der Erstellung beliebiger kritischer, historischer bzw. genetischer Editionen.

Weitere Funktionen, die das Spektrum von eComparatio noch einmal entscheidend erweitern werden, sind in Entwicklung. So ist die Einbindung von hochauflösenden Images der Handschriften der betreffenden Editionen geplant, um auch diesen Abschnitt der Textgeschichte dem Nutzer zugänglich zu machen. Weiterhin ist ein weiteres Modul in Entwicklung, das für die Abfrage von online-Datenbanken die Anbindung an das Notationssystem CTS (Canonical Text Services) ermöglicht. Beide Erweiterungen des Tools werden in absehbarer Zeit implementiert werden.

Nach seiner Fertigstellung soll das Tool als freier Webservice für Forschung und Lehre zur Verfügung gestellt werden. Davon können Handschriften-Digitalisierungsprojekte, Editionsprojekte sowie Projekte profitieren, die sich Spezialfragen einzelner Textpassagen widmen; es ist auch für Seminararbeiten, d.h. den Einsatz in der Lehre geeignet, da es sowohl von Nicht-Editionsphilologen als auch von Editionsphilologen eingesetzt werden kann. Es ist natürlich auch nicht an den Fachbereich der Alten Geschichte gebunden, sondern kann in verschiedenen Bereichen der Textwissenschaften, unabhängig von der Sprache, eingesetzt werden.

In der Fachcommunity der E-Humanities im Speziellen kann das Tool darüber hinaus in einem Bereich angewandt werden, der in jüngerer Zeit vermehrt ins Zentrum der Aufmerksamkeit gerückt ist, nämlich bei der Qualitätssicherung der digitalen Datengrundlage an sich. Gerade im Falle der Altertumswissenschaften, in denen bereits früh umfangreiche, abgeschlossene Korpora (TLG, BTL u.a.) vorlagen, ist ein nächster Schritt ein Ausbau dieser Datengrundlagen in die Tiefe, d.h. hinsichtlich der zahlreichen verschiedenen Editionen und Textausgaben. Solche Varianten spielen in der herkömmlichen altertumswissenschaftlichen Diskussion oftmals eine zentrale Rolle bei der Erörterung fachwissenschaftlicher Fragestellungen; die Möglichkeit, solche Varianten im Falle auch großer Textmengen schnell zu überblicken, kann als eine wesentliche Grundlage dafür gesehen werden, auch auf „klassischem“ Textmining basierende Untersuchungen mit einer besseren Datengrundlage zu versehen.

Da es sich bei dem Tool in erster Linie um ein Mittel zur Visualisierung handelt, ist es in hohem Maße für die Präsentation in Form eines Posters geeignet. Geplant ist die Darstellung des gesamten Workflows anhand eines Beispiels, von der Eingabe unstrukturierter Textdokumente bis hin zu den drei oben genannten Visualisierungsformen.

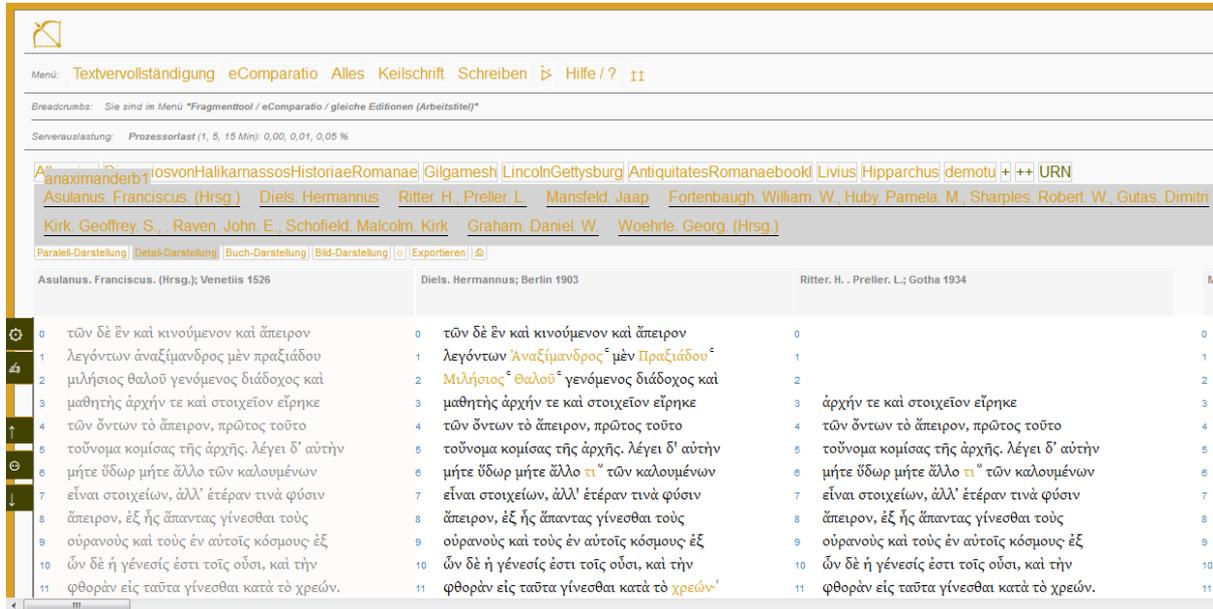


Abb. der Parallelansicht von eComparatio am Beispiel des Fragments B1 des Anaximander.

Kontakt:

Prof. Dr. Charlotte Schubert

Historisches Seminar

Lehrstuhl für Alte Geschichte

Beethovenstraße 15

04107 Leipzig

Raum 3.204

Telefon: +49 341 97 37071

Email: schubert@rz.uni-leipzig.de

»So viele Briefe mit all ihrem Für und Wider...« Die kommentierte Online-Edition des Gesamtbriefwechsels Ludwig von Fickers als wissenschaftlicher Quellenfundus

Markus Ender
Forschungsinstitut Brenner-Archiv
Universität Innsbruck

Ludwig von Ficker (1880–1967) erlangte als Entdecker und Förderer Georg Trakls und als Herausgeber der Zeitschrift »Der Brenner« (1910–1954) einige Bekanntheit; daneben betätigte er sich als Inhaber des Brenner-Verlags, als Literaturkritiker, Juror und Organisator von Lesungen. Aufgrund seiner vielen Tätigkeiten ergaben sich enge briefliche Kontakte mit Personen aus Politik und Kultur, so z.B. mit Else Lasker-Schüler, Martin Heidegger, Karl Kraus, Rainer Maria Rilke oder Ludwig Wittgenstein. Sein Briefwechsel, von dem im Innsbrucker Forschungsinstitut Brenner-Archiv mehr als 16.500 Korrespondenzstücke von über 2200 AdressatInnen erhalten sind, markiert und dokumentiert einen Teil der deutschsprachigen Kulturgeschichte und bietet Forscherinnen und Forschern wie auch interessierten Laien Einblicke in das Geistesleben der ersten Hälfte des 20. Jahrhunderts. Zwischen 1988 und 1996 erschien in vier Bänden eine Auswahl von 1300 Briefen von und an den »Brenner«-Herausgeber.¹

Im Rahmen des FWF-Projektes »Ludwig von Ficker als Kulturvermittler« (P24283) entsteht seit April 2012 am Brenner-Archiv eine digitale Ausgabe des Briefwechsels Ludwig von Fickers in Form einer kommentierten Online-Edition. Diese Edition versteht sich nicht als bloße Retrokonversion der bereits gedruckten, vierbändigen Auswahlausgabe von Fickers Briefen, sondern als eigenständige Neu-Edition, die in ihrer Konzeption, der methodischen Durchführung, in ihrer Darstellungsform und in ihrem intendierten Gebrauchswert in wesentlichen Punkten von der früheren Buchausgabe abweichen wird.

Die digitale Edition der Fickerschen Korrespondenz wird dabei in mehrfacher Hinsicht als eine integrale Schnittstelle zwischen den Beständen im Brenner-Archiv und den RezipientInnen dienen; im Sinne des Tagungsthemas möchte ich in meinem Beitrag die Leistungsfähigkeit einer solchen Editionsform aufzeigen. Es soll am Beispiel der kommentierten Online-Edition des Briefwechsels Ludwig von Fickers demonstriert werden, dass sie sich sowohl als Medium für die zukünftige Generierung von Wissen als auch für die nachhaltige Nutzung von Daten eignen kann. Diesbezüglich lassen sich drei große Bereiche anführen:

- Zum einen bietet die Form der digitalen Internet-Edition auf *quantitativer* Ebene die Möglichkeit, erstmalig den gesamten Archivbestand im Nachlass Ludwig von Fickers (also Briefe und Gegenbriefe) zugänglich zu machen.² Die Online-Edition betrachtet die vorliegenden Briefwechsel (im Sinne Foucaults) wertfrei als Summe von Aussagen, die zu einem bestimmten Zeitpunkt möglich waren, wobei nicht zwischen »wichtigen« und »unwichtigen« BriefpartnerInnen bzw. Briefen unterschieden wird. Die Breite der vorliegenden Daten ermöglicht einen neuen Blick auf das Gesamtkorpus, der in dieser Form bislang nicht möglich gewesen ist.
- Zum anderen spricht ein solch umfangreiches Briefkonvolut, das einen gewichtigen Baustein im kulturellen Erbe darstellt, auf *qualitativer* Ebene durch die Veröffentlichung im Rahmen eines methodisch kontrollierten und dokumentierten

¹ Ludwig von Ficker: Briefe. 4 Bde. Hg. von Franz Seyr, Walter Methlagl u. a. Salzburg; Innsbruck 1988–1996.

² Diesem Ansatz wird insofern Rechnung getragen, als dass bereits 14000 Briefe als Transkripte vorliegen.

Editionsprojekts ein breites Spektrum von InteressentInnen an. Die Daten werden über das Internet sowohl einer interessierten Öffentlichkeit als auch der wissenschaftlichen Forschung zugänglich gemacht; der Briefwechsel dürfte dabei aufgrund der inhaltlichen Diversität, der erschließenden Kommentierung sowie der Möglichkeit, über spezifizierte Suchfunktionen personelle und thematische Netzwerkstrukturen auszumachen, nicht nur für Literaturwissenschaftler, sondern für verschiedene Fachrichtungen (so z.B. Soziologie, Geschichtswissenschaften oder Theologie) von beträchtlichem Interesse sein.

- Überdies kommen bei der Erstellung der Edition etablierte digitale Standards (so z. B. das XML-Dateiformat oder die TEI-Codierung) zur Anwendung. Dadurch ist in Folge auch für die Institution Archiv ein Zusatznutzen gewährleistet, denn es kann durch die Digitalisierung der Korrespondenz eine seiner Hauptaufgaben, die nachhaltige Langzeitarchivierung der Bestände, wahrnehmen. Die den Transkripten zugrunde liegende XML-Datenstruktur garantiert bei zukünftigen Bearbeitungen die volle Verfügbarkeit der editorischen Kerndaten (Brieftranskripte) sowie der darauf aufbauenden Metadaten (inklusive Kommentar etc.). Die dem Editionsprojekt zugrundeliegende Open-Access-Policy und die geplante Anbindungen an die vom Austrian Academy Corpus besorgte Online-Version des »Brenner« sowie an die Gemeinsame Normdatei der Deutschen Nationalbibliothek erweitern das mögliche Nutzungsspektrum der Edition.

Kontakt:

Mag. Markus Ender
Forschungsinstitut Brenner-Archiv
Universität Innsbruck
Josef-Hirn-Straße 5-7
A-6020 Innsbruck
Tel. +43 512 507 45022
E-Mail: Markus.Ender@uibk.ac.at
<http://www.uibk.ac.at/brenner-archiv/projekte/lfickeralskulturvermittler/>

Geometrische Verfahren als Brücke zwischen Text und Objekt

Hubert Mara und Bartosz Bogacz

Universität Heidelberg
IWR – Interdisziplinäres Zentrum für Wissenschaftliches Rechnen
FCGL - Forensic Computational Geometry Laboratory
Im Neuenheimer Feld 368, 69120 Heidelberg, Deutschland
hubert.mara@iwr.uni-heidelberg.de

Keilschrifttafeln gehören zu den ältesten Textzeugen, die im Umfang mit den Texten in lateinischer und altgriechischer Sprache vergleichbar sind. Da diese Tafeln aus dem gesamten Alten Orient über beinahe viertausend Jahre in Verwendung waren [Sod94], lassen sich damit viele interessante Fragestellungen zur Entwicklung von Religion, Politik, Wissenschaft, Handel bis hin zu Klimaveränderungen [Kan13] beantworten. Die aus Ton geformten Tafeln, bei denen Zeichen [Bor10] als keilförmige Abdrücke mit einem eckigen Stylus eingedrückt wurden, erfordern neue informationstechnische Methoden zu der Dokumentation und Analyse als die in Archiven üblichen Flachwaren. Zusätzlich gibt es kaum Verfahren aus dem Bereich der *Optical Character Recognition* (OCR), die für Sprachen in Keilschrift zur Verfügung stehen [Spe81]. Der Arbeitsablauf in der Assyriologie von der Keilschrifttafel als dreidimensionales Objekt bis hin zur Darstellung als Text in einer modernen Sprache – üblicherweise Deutsch – ist in Abbildung 1 dargestellt. Dabei ist der manuelle Zeitaufwand als Kurve dargestellt. In der Zusammenarbeit mit der Heidelberger Assur-Forschungsstelle unter der Leitung von Prof. Stefan Maul konnte festgestellt werden, dass ein erheblicher Teil der Arbeit im Bereich der Identifikation und Extraktion von Zeichen liegt, der stark mit der Dokumentation als Zeichnung verknüpft ist.

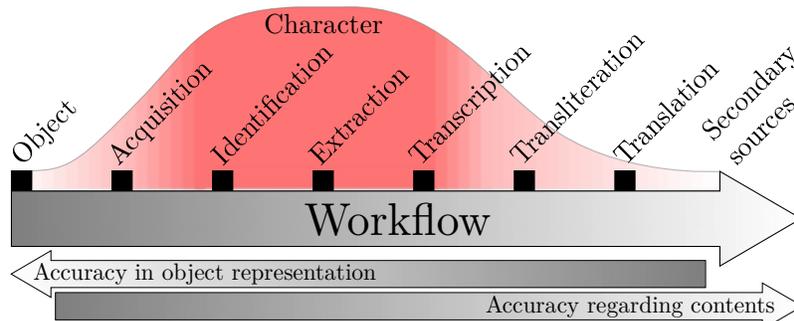


Abb. 1: Arbeitsschritte von der Keilschrifttafel als Objekte bis zur Übersetzung als Text in einer modernen Sprache.

Das Digitalisieren von Keilschrifttafeln inspiriert durch die Open Data Initiative, wurde bereits vor einigen Jahren von der *Cuneiform Database Library Initiative* (CDLI) des Max Planck Institut für Wissenschaftsgeschichte und der *University of California at Los Angeles* begonnen und entsprechende Datenbanken entwickelt [GWL05]. Dabei werden üblicherweise Photos und Bilder von Flachbettscannern eingesetzt, die zwar günstig und rasch zu erstellen sind, jedoch bei beschädigten oder gekrümmten Tafeln viele Bereiche unscharf und/oder verschattet sind. Daher werden in Jena, Würzburg [CMFW14] und Heidelberg [MKJB10] moderne 3D-Messgeräte eingesetzt um möglichst exakte digitale Repliken anzufertigen, mit denen entsprechende Visualisierungen berechnet werden. Als Mittel- und Fernziel sind digitale Werkzeuge im Sinne der OCR in Entwicklung.

Da die Datengrundlage keine regelmäßigen Gitter i.e. Rasterbilder wie die in *Digital Humanities* üblichen 2D-Digitalisate sind, sind Methoden notwendig, die aus der Geometrie eines 3D-Modells die Schriftzeichen extrahiert. Dafür kommen Integral Invariante Filter zum Einsatz, die mit Hilfe eines Mehr-Skalen Ansatzes die einzelnen Elemente der Keilschriftzeichen in einer Vektordarstellung extrahieren [MK13]. Mit Hilfe von einer minimalen Anzahl von Orts- und Richtungsvektoren werden mit dem *GigaMesh Software Framework* parametrischen Kurven (i.e. Splines) bestimmt, die als zweidimensionale XML-Dateien im offenen *Scalable*

Vector Graphics (SVG) Format exportiert werden. In Abbildung 3 wird eine schematisches SVG zur Darstellung eines Keils, mit einer minimalen Menge von vier Punkten repräsentiert wird.

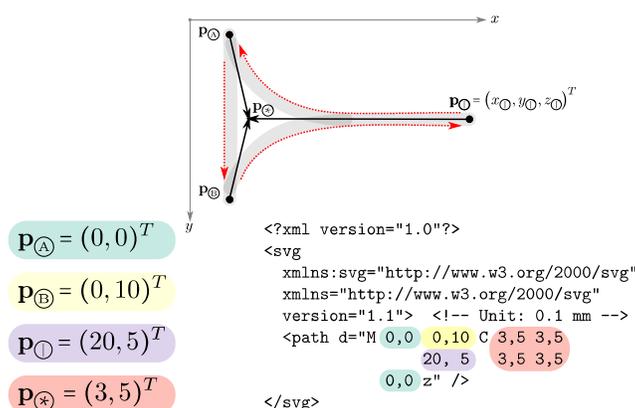


Abb. 2: Minimalbeispiel für einen Keil als XML/SVG-Datei.

Das selbe Format wird von proprietären Zeichenprogrammen und dem *Open Source* Pendant Inkscape verwendet, die beide in der Assyriologie und in der Grabungsdokumentation in der Archäologie zum Einsatz kommen. Damit ist automatisch sichergestellt, dass aus 3D-Modellen berechnete Zeichnungen kompatibel sind zu digitalen Handzeichnungen. Darüber hinaus bietet SVG – wie alle anderen – XML-Dateien die Möglichkeit zur automatischen und manuellen Annotation, wie es in den digitalen Textwissenschaften üblich ist. Abbildung 3 zeigt einen Vergleich zwischen einer digitalen Handzeichnung und einer berechneten Zeichnungen des zugehörigen 3D-Modells.

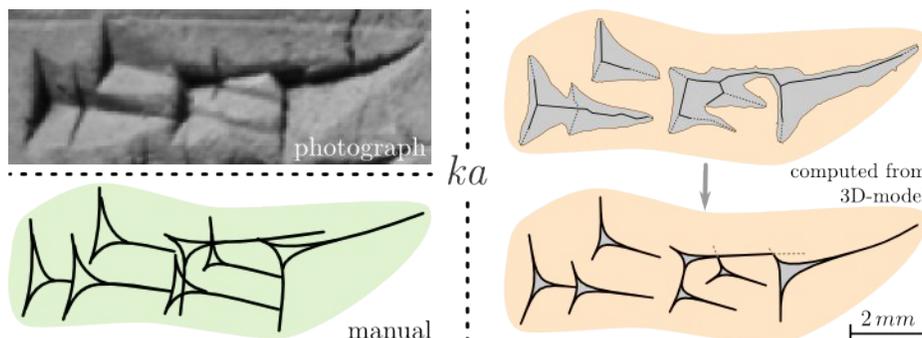


Abb 3: Das Zeichen „ka“ digital mit der Hand gezeichnet und aus einem 3D-Modell berechnet.

Die Vektordarstellung der Keilschriftzeichen sowie die komplexe zweidimensionale Anordnung von Keilabdrücken verhindert die Anwendung von gebräuchlichen OCR Methoden, die Zeichen in Rasterdarstellung [AGFV14] und aufeinander folgende Zeichen [RRF13] erwarten. Die Analyse von Keilschriftzeichen erfordert eine Transformation der SVG Daten in eine vereinfachte aber mathematisch handhabbare Repräsentation als mathematische Graphen mit Knoten und Kanten. Keilschriftzeichen identifizieren sich hauptsächlich durch die Lage und Position ihrer Keilabdrücke, eine Eigenschaft, die sich mit der Zerlegung des Graphen in Teilstrukturen, die den Keilabdrücken entsprechen, nutzen lässt. Der Teilabdruck als Teilstruktur in einem Graphen lässt sich einfach mit Richtungs- und Ortsvektoren beschreiben, die als Features genutzt werden, um Keilschriftzeichen auf Ähnlichkeit zu prüfen. Die vollständigen Graphen der Zeichen werden zudem genutzt, um Methoden aus dem Gebiet der Graphenähnlichkeit, wie den Graph-Kernen und dem spektralen Embedding [BR10] anzuwenden. Dies ist vor allem vorteilhaft bei komplexen bildhaften Zeichen, die sich nicht in Teilstrukturen von Features zerlegen lassen, aber trotzdem auf Ähnlichkeit verglichen werden müssen.

Abbildung 4 zeigt Keilschriftzeichen dargestellt als Graphen, die aus einer SVG Datei extrahiert wurden, und gegenseitig auf Ähnlichkeit verglichen werden. Die Aufgabe besteht darin Keilschriftzeichen, die einem Zeichen (L: 19) ähneln, aufzufinden. Das erste Zeichen von Links ist der gesuchte Prototyp, alle darauf folgende Zeichen

sind die gefundenen Zeichen. Die Zeichen mit grünem Hintergrund wurden korrekt identifiziert, die Unähnlichkeit zum Prototyp wird mit „k:“ unterhalb des Zeichens betitelt. Alle fünf Zeichen, die in dem analysierten Dokument vorhanden waren und dem gesuchten Prototyp ähnelten, wurden erfolgreich gefunden.

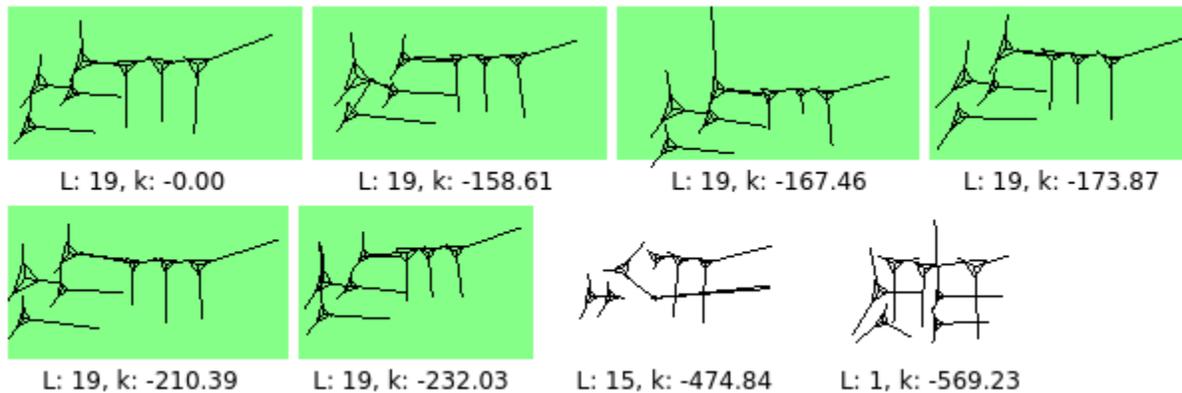


Abb. 4: Keilschriftzeichen in Graphenrepräsentation werden in einem SVG Dokument gesucht und auf Ähnlichkeit verglichen.

Zusammenfassend werden in diesem Beitrag Methoden aus der Geometrie und der Mustererkennung vorgestellt, die frei von lexikographischen / linguistischen Annahmen sind. Damit werden neue Zugänge zur Integration in OCR System für Handschriften geschaffen, die weit über die Anwendung an Keilschrift i.e. Handschrift in 3D hinaus gehen. Eine direkte Anwendung an mittelalterlichen Epitaphen hat bereits Aufnahme in entsprechende online Datenbanken gefunden [Krö12].

Literatur

- [AGFV14] Segmentation-free word spotting with exemplar SVMs, J. Almazán, A. Gordo, A. Fornés, E. Valveny, *Journal of Pattern Recognition*, pp. 3967-3978, Elsevier, 2014.
- [Bor10] R. Borger. Mesopotamisches Zeichenlexikon, volume 305 of *Alter Orient und Altes Testament – Veröffentlichungen zur Kultur und Geschichte des Alten Orients und des Alten Testaments (AOAT)*. Ugarit-Verlag, 2. edition, 2010.
- [BR10] Recent advances in graph-based pattern recognition with applications in document analysis, H. Bunke, K. Riesen, *Journal of Pattern Recognition*, pp. 1057-1067, Elsevier, 2010.
- [CMFW14] M. Cammarosano, G. G.W. Müller, D. Fisseler and F. Weichert. *Schriftmetrologie des Keils: Dreidimensionale Analyse von Keileindrücken und Handschriften*, *Die Welt des Orients*, [Ausgabe: 44.1](#), 2014.
- [GWL05] B. Groneberg, F. Weiersh#user, T. Linnemann, and D. Ullrich. *Jahrbuch der Max-Planck-Gesellschaft, chapter Digitale Keilschriftbibliothek Lexikalischer Listen aus Assur*. Gesellschaft für wissenschaftliche Datenverarbeitung mbH, Göttingen, Germany, 2005.
- [Kan13] D. Kaniewski, E. Van Campo, J. Guiot, S. Le Burel, T. Otto and C. Baeteman. Environmental Roots of the Late Bronze Age Crisis. *PLoS ONE* 8(8), 2013.
- [Krö12] S. Krömker, Kombinierte 3D-Datenaufbereitung von Schriftfeldern und Gelände des mittelalterlichen Jüdischen Friedhofs ‚Heiliger Sand‘, in: *Die SchUM-Gemeinden Speyer–Worms–Mainz. Auf dem Weg zum Welterbe. Band zur Internationalen Tagung der Generaldirektion Kulturelles Erbe Rheinland-Pfalz*, angenommen, Mainz, Deutschland, 2012.
- [MK13] H. Mara and S. Krömker. Vectorization of 3D-Characters by Integral Invariant Filtering of High-Resolution Triangular Meshes. *Proc. of 12. Int. Conference on Document Analysis and Recognition (ICDAR/IAPR)*, pp. 62–66, Washington, DC, USA, 2013.
- [MKJB10] H. Mara, S. Krömker, S. Jakob and B. Breuckmann. GigaMesh and Gilgamesh - 3D Multiscale Integral Invariant Cuneiform Character Extraction. *Proc. VAST Int. Symposium on Virtual Reality, Archaeology and Cultural Heritage*, pp. 131-138, Palais du Louvre, Paris, France, 2010.
- [RRF13] Bag-of-Features HMMs for segmentation-free word spotting in handwritten documents, L. Rothacker, M. Rusinol, G.A. Fink, *Proc. of 12th International Conference on Document Analysis and Recognition*, pp. 1305-1309, Washington, DC, USA, 2013.
- [Sod94] W. von Soden. *The ancient Orient: an introduction to the study of the ancient Near East*. Wm. B. Eerdmans Publishing Co., 1994.
- [Spe81] G. Sperl. *Erkennen von Keilschriftzeichen mit Hilfe Elektronischer Rechenanlagen*. PhD thesis, Leopold-Franzens-Universität Innsbruck, Innsbruck, Austria, 1981.

Vernetzte Datenstrukturen als Grundlage philosophischer Erkenntnisse
Technische Umsetzung und eine exemplarische Anwendung anhand des elektronischen
Apparats der WIENER AUSGABE

Michael Nedo, Daniel Bruder, Pascal Zambito, Max Hadersbeck, Josef Rothhaupt

The Wittgenstein Project Clare Hall und der Ludwig Wittgenstein Trust, University of
Cambridge

Centrum für Informations- und Sprachverarbeitung, in Zusammenarbeit mit dem
Lehrstuhl II der Fakultät für Philosophie, Wissenschaftstheorie und Religionswissenschaft,
LMU München

1. Einleitung

Ludwig Wittgenstein, einer der bedeutendsten Denker und Philosophen unserer Zeit, eignet sich aufgrund seiner Themen und seiner komplexen Arbeitsweise besonders gut, um die Vorteile eines elektronischen Apparats auf Grundlage vernetzter Datenstrukturen zu demonstrieren. Da bisher publizierte Editionen ohne solche Werkzeuge auskommen mussten, lassen sich durch geschickte Nutzung der digitalen Möglichkeiten neuartige philosophische Erkenntnisse gewinnen und alte Irrtümer ausräumen.

Auf unserem Poster wollen wir die zugrunde liegenden Datenstrukturen des Apparats erklären und in einer exemplarischen Anwendung zeigen, wie sich konkreter Nutzen aus seiner Anwendung ziehen lässt

2. Struktur von Wittgensteins Werk

Wittgensteins Werk zeichnet sich durch eine Vielzahl an internen und externen Verknüpfungen aus. Das Stemma in Abb. 1 zeigt u.a. die Entstehung der bei Suhrkamp publizierten *Philosophischen Bemerkungen (PB)*, an der deutlich wird, dass die Genese des Textes sowohl editionsphilologisch als auch philosophisch relevant ist.

Die Entstehungsgeschichte der *PB* liefert Aufschluss darüber, wie Bemerkungen aus den Manuskripten Eingang in die maschinenschriftliche Synopse TS208 gefunden haben. Die Synopse wurde von Wittgenstein zerschnitten, in der Zettelsammlung (TS209) neu angeordnet und schließlich unter Hinzuziehung weiterer Materials aus anderen Manuskripten von seinen Erben als Buch veröffentlicht.

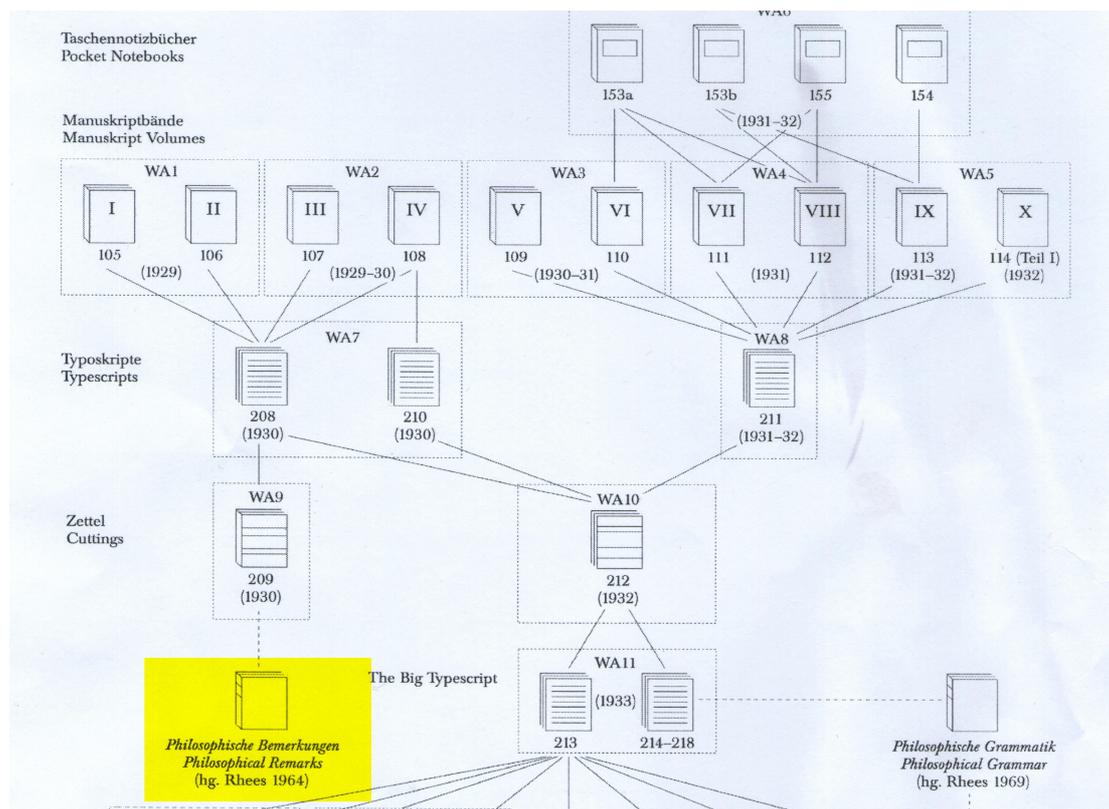


Abbildung 1: Stemma zur Genese der *Philosophischen Bemerkungen*

Da die meisten Interpreten von diesem Werk als fertiges Buch ausgingen, finden sich bei ihnen Missverständnisse, die durch eine angemessene Darstellung der intra- und intertextuellen Verknüpfungen vermeidbar sind. Der Apparat, an dem wir arbeiten, ist ein digitales Werkzeug, das mittels verschiedener Such- und Sortierungsfunktionen diese Interdependenzen dem Nutzer verständlich macht und ihm einen ebenso einfach zu handhabenden wie zuverlässigen Zugang zum Werk erlaubt.

3 Datenstruktur des Apparats

Die Nutzung des elektronischen Apparats beginnt in der Regel mit einer Wortsuche, welche über eine lemmatisierte Wortkonkordanz realisiert wird. Den weiteren Funktionen - Zugriff auf eine elektronische Realkonkordanz und interaktive Nutzung durch den User - liegt eine Matrix dynamisch vernetzter Objekte zugrunde.

3.1 Wortkonkordanz

Mithilfe des Computers lassen sich Suchfunktionen effizienter und benutzerfreundlicher realisieren als in den gedruckten Apparaten zur WIENER AUSGABE. Zur Wortsuche wird ein Vollformenlexikon genutzt, welches es ermöglicht, aus jeder flektierten Form eines Wortes eine lemmatisierte Form

abzuleiten und daraus wiederum sämtliche möglichen Formen des Wortparadigmas zu generieren und in die Suche miteinzuschließen. Indem das Suchergebnis also nicht nur wortidentische Treffer, sondern alle linguistisch verwandten Formen enthält, erlaubt es die Wortkonkordanz, das gesamte Korpus auf ein spezifisches Thema oder einen bestimmten Begriff hin „quer“ zu lesen.

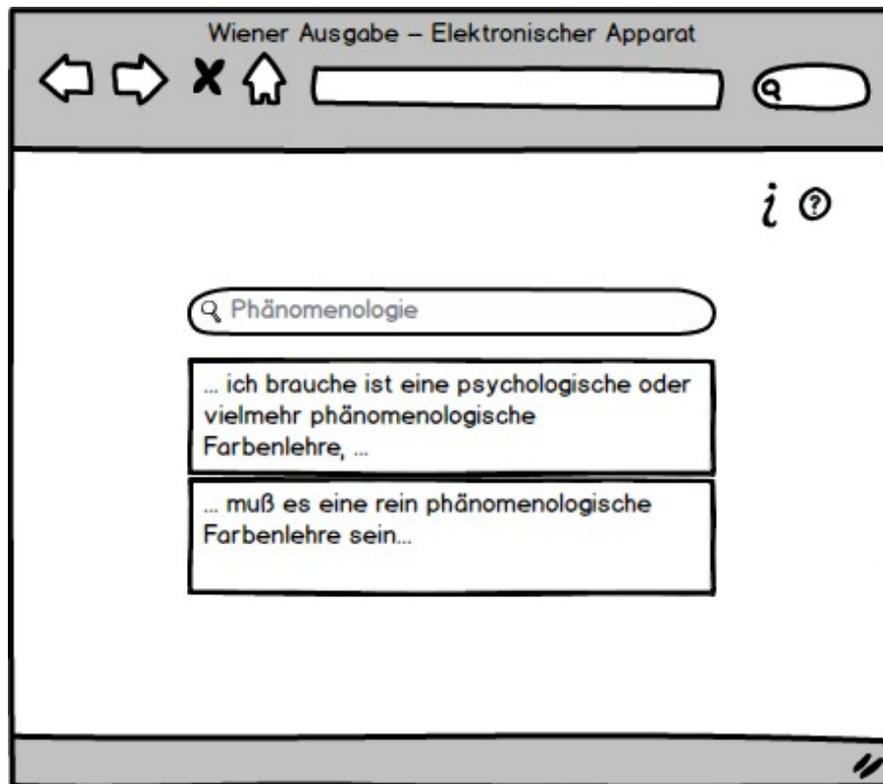


Abbildung 2: Eingabemaske für die Wortsuche

3.2 Realkonkordanz

Die Funktionsstruktur der Realkonkordanz basiert auf einer mit dem Werk dynamisch vernetzten Matrix. Auf der ersten Ebene werden die Objekte, die das Werk repräsentieren, aufgeführt: Wittgensteins Manuskripte, Typoskripte und Mitschriften sowie Diktate und deren jeweilige Veröffentlichungen. Die internen Strukturen der Objekte entsprechen der Nomenklatur der WIENER AUSGABE (vgl. Abb. 3, A); bei den posthumen Veröffentlichungen werden stattdessen die Strukturen der Herausgeber aufgeführt.

Entsprechend dieser Grundstruktur werden auf separaten Ebenen weitere Objekte erfasst, wie z.B. biographische Dokumente, Korrespondenzen, Bilder oder auch Sekundärliteratur und Übersetzungen (B).

Zusätzlich werden über eine interaktive Arbeitsplattform unter den Namen der jeweiligen Nutzer inhaltsbezogene Kommentare mit den Texten verknüpft sowie Fehler und vorgeschlagene Korrekturen in den Editionen (D).

Die Vernetzung der einzelnen Elemente der Objekte erfolgt auf darunter liegenden Ebenen der Matrix. Dabei werden drei Typen von Verknüpfungen unterschieden: ein-eindeutige, quasi festverdrahtete; provisorische, noch endgültig zu bestimmende; und offene, die noch recherchiert werden müssen (C).

Die Matrix ist über Such- und Sortierungswerkzeuge mit den Textdateien und deren Darstellungsformen auf dem Computerbildschirm verknüpft, die wiederum über die Nomenklatur der WIENER AUSGABE auf die gedruckte Edition verweisen.

WERKIMMANENT (A)

- Objektnamen
- Seitennummern
- Bemerkungsnummern
- Absatznummer
- Randzeichen

AUSSENBEZUG (B)

- Sekundärliteratur
- Übersetzungen
- Biographische Anm.
- Facsimile

VERKNÜPFUNGEN (C)

- Feste, eindeutige
- Provisorische
- Offene

INTERN (D)

- Darstellung der Seite
- Textbausteine
- Benutzerverwaltung
- Benutzerkommentare
- etc.

4. Anwendung

Abbildung 3: Datenstrukturen der Matrix

Im Rahmen einer Masterarbeit an der LMU werden diese Datenstrukturen genutzt, um philosophische Erkenntnisse zu gewinnen. Konkret geht es um das Themenfeld der Phänomenologie bei Wittgenstein in den Jahren 1929-30, in denen der Begriff signifikante Bedeutungsveränderungen erfuhr. Mittels der Wortkonkordanz lassen sich zunächst alle Vorkommen des Begriffs im entsprechend zeitlich eingegrenzten Korpus finden (vgl. Abb. 2). Die einmal gefundene Textstelle lässt sich dann

1. im Rahmen des betreffenden Objektes um den Kontext erweitern, sodass benachbarte Stellen etwa in TS209 angezeigt werden
2. um frühere oder spätere Versionen der gleichen Bemerkung erweitern. Bemerkungen aus den TS209 lassen sich z.B. rückwärts bis zu den Manuskriptbänden oder vorwärts bis zur Publikation in den *PB* nachverfolgen. (s. Abb. 1)
3. um Hintergrundinformationen erweitern, die zu jeder Textstelle Faksimiles, Sekundärliteratur und weitere Daten enthalten können.

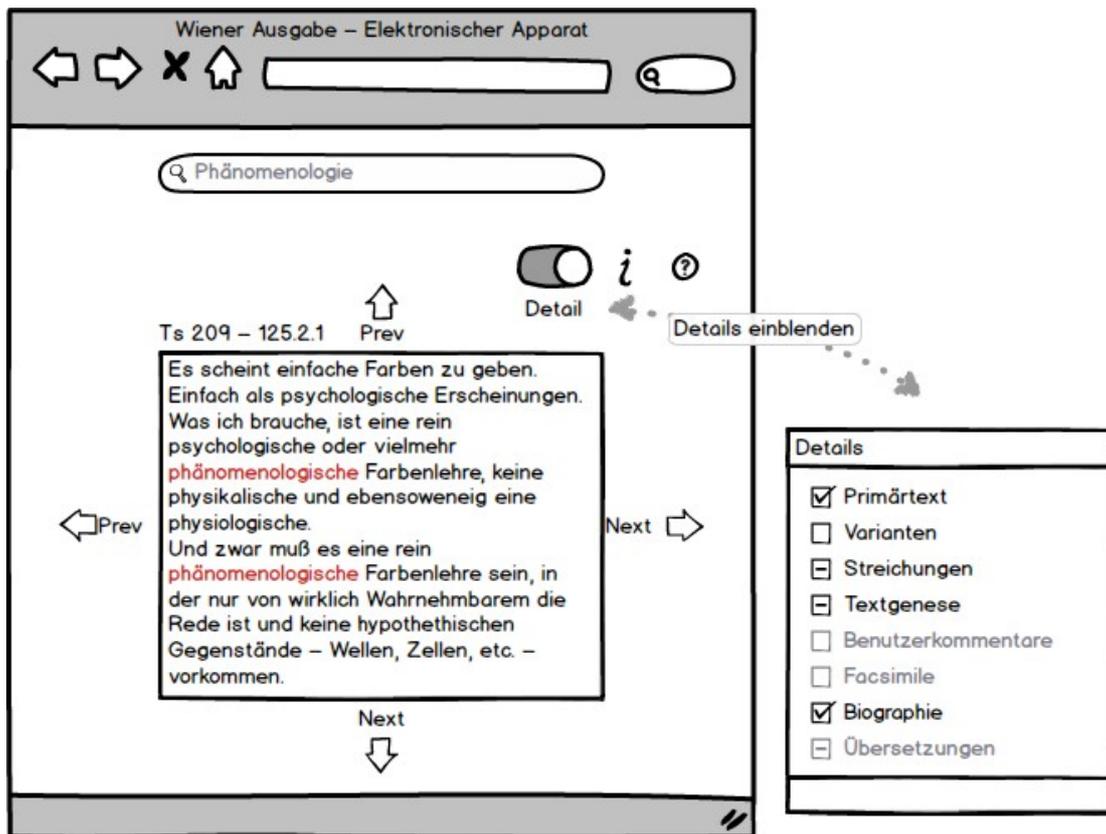


Abbildung 4: Die Bemerkung TS209,125.2.1/2.2 in der Darstellung des Apparats

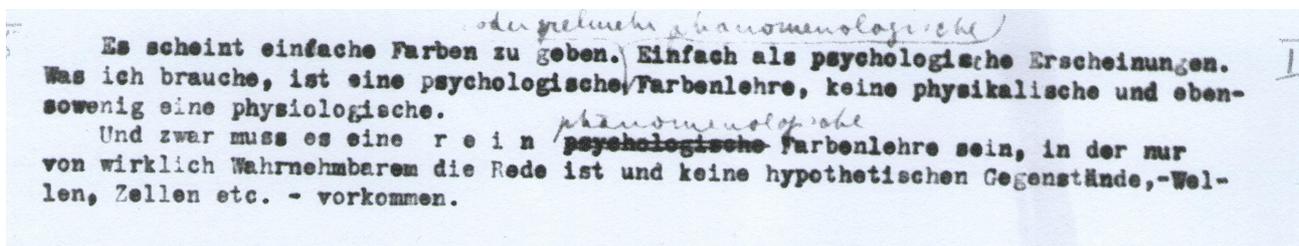


Abbildung 5: Faksimile derselben Bemerkung: man sieht, dass das Wort "phänomenologische" erst nachträglich eingefügt wurde und könnte nun frühere Textstufen untersuchen, um die Begriffsentwicklung zu erforschen.

Auf Grundlage dieses Wissens lassen sich gängige Missverständnisse in der Sekundärliteratur ausräumen, der die Genese der PB nicht in diesem Ausmaß zugänglich war. Schließlich lassen sich die neu gewonnenen Erkenntnisse verknüpft mit dem Namen des Verfassers als Kommentar zurück in die Matrix einspeisen.

5. Anhang- Technische Details

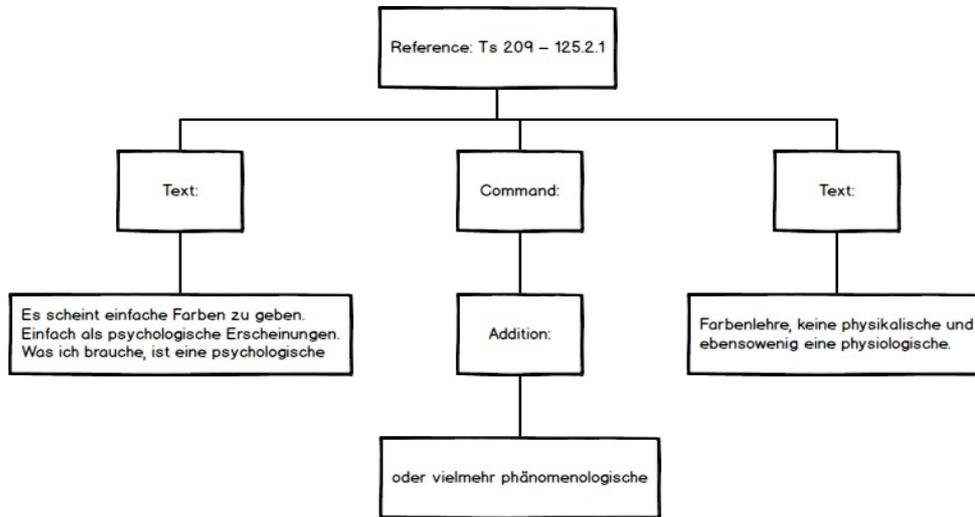


Abbildung 6: Abschnitt eines Abstract Syntax Tree (AST) für Bem. TS209, 125.2.1

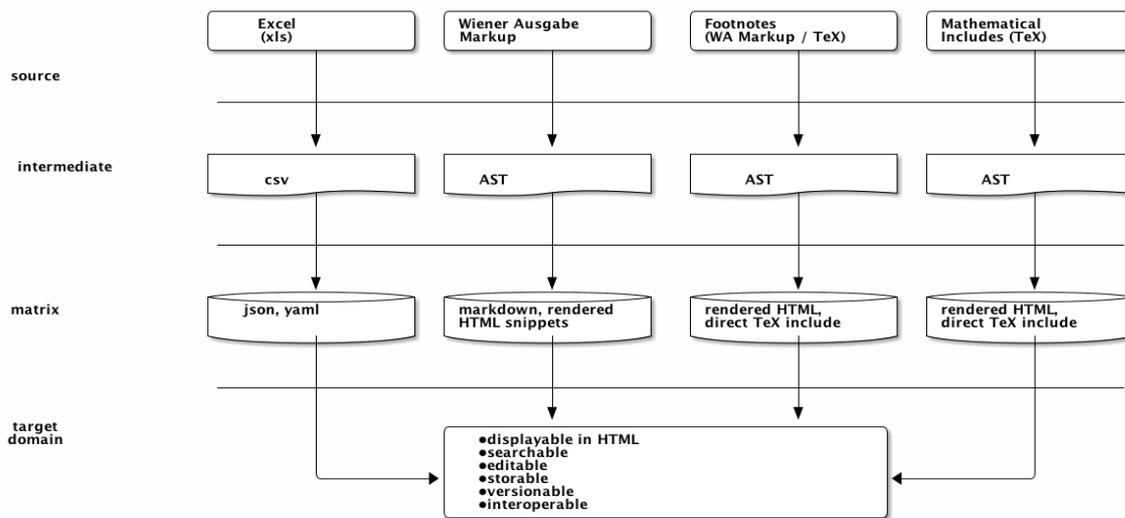


Abbildung 7: Herstellung und Funktionsweise von AST

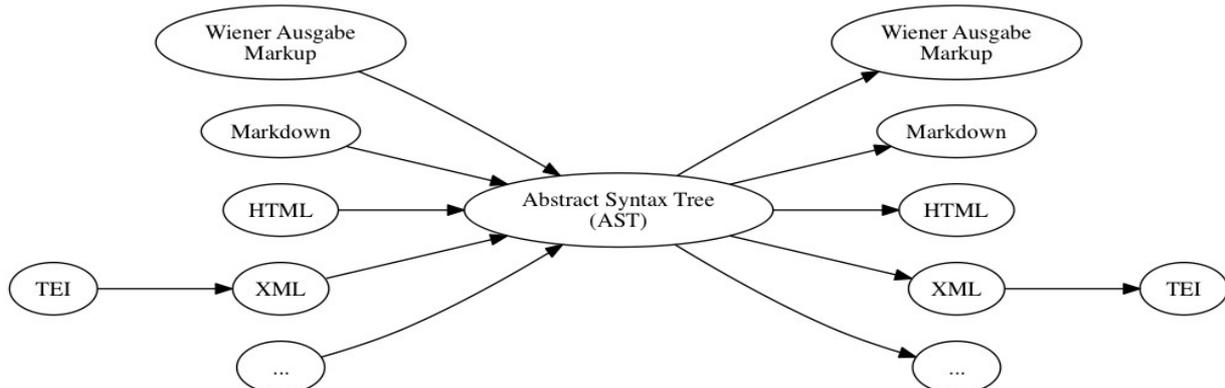


Abbildung 8: Herstellung und Funktionsweise von AST

Poster

Visualisierung von Kultur im Web

Das Poster problematisiert die softwaregestützte Visualisierung von Kultur im World Wide Web (Web). Bisher konzentrieren sich sozialwissenschaftlich relevante Visualisierungen auf das Web als Netzwerk von Informationen. Untersucht wird vor allem der menschliche Faktor in Verbindung mit Hypertext und Internetinfrastrukturen bei der Entstehung, Verbreitung und Veränderung von verknüpften Informationen und Informationsflüssen. Dazu werden in der Regel Beziehungen, Verteilungen und die Performance auf der Basis von Metadaten durch vereinfachte visuelle Darstellungen (Punkte, Linien, einfache geometrische Figuren etc.) in Häufigkeitsverteilungen, Netzwerkbeziehungen oder Pfadmodellen abgebildet. Eindrucksvolle Studien visualisieren beispielsweise Beziehungsmuster zwischen den NutzerInnen in Facebook oder die NutzerInnen-Performance in Wikipedia.

An solchen Visualisierungen kann mit Manovich (2011) die Reduktionen der Daten auf ein Merkmal (die oben angesprochenen vereinfachten Darstellungsformen) oder die Präferenz, Eigenschaften vor allem räumlich zu visualisieren (z.B. nah/fern, innen/außen), kritisiert werden. Ebenso kann man die dominante Netzwerklogik bei der Beobachtung des Webs problematisieren. Die einseitig erfolgte Zuspitzung ist jedoch legitim, um die Vernetzung von Informationen zu repräsentieren und zu beobachten. Zugleich besteht das Web nicht nur aus verknüpften Informationen, sondern es erzeugt auch spezifische Bedeutungszuschreibungen und kollektive Deutungsmuster. Es handelt sich folglich, kultursoziologisch ausgedrückt, beim Web ebenfalls um ein von Menschen geschaffenes Bedeutungsgewebe (Geertz 1987). Mit Gewebe ist aber eben kein Netzwerk aus Informationen gemeint, sondern kulturspezifische Auslegungen und Deutungen (kurz: Sinnzusammenhänge, Frames, Orientierungsschemata). Für die textbasierte Untersuchung und Analyse von Strukturen und Mustern bieten sich in der Regel direkte Visualisierungen wie Tag Clouds, statistisch gestützte Kookkurrenzanalysen oder Clusterverfahren wie das Topic Modeling an. Für die Analyse von Kultur liefern diese Verfahren Möglichkeiten, thematische Differenzierungen und die Relevanz bestimmter Themen punktuell über die Zeit abzubilden. Insbesondere für online zugängliche Pressemitteilungen und Pressenachrichten liegen dazu bereits verschiedene, überzeugende Studien vor (z. B. Mohr & Bogdanov 2013).

Textproduktionen in Organisationen wie Unternehmen, Medienanstalten oder Nicht-Regierungsorganisationen unterliegen aber häufig hierarchischen Entscheidungsstrukturen. In der Regel existieren Formen institutionalisierter Autorisierung, die über die Herstellung, die

Inhalte und Freigaben von Texten entscheiden. Das Web bietet hingegen auch die Möglichkeit, kulturspezifische Deutungen und Auslegungen jenseits institutioneller Autorisierungen zu beobachten. Zu diesen Sinnzusammenhängen zählen „folksonomies“ (Mathes 2004), welche vor allem in den sozialen Medien entstehen. Derartige „grassroots“-Kategorisierungen in Form von „tags“ (nutzergenerierte Schlagworte), um beispielsweise digitale Medien wie URLs oder Photographien zu teilen und zu organisieren, entsprechen alltagstheoretischen, kommunikativen Typisierungen. Aus einer kultursoziologischen Perspektive bilden solche Typisierungen ein erfahrungsbasiertes Orientierungswissen, welches die alltägliche Praxis strukturiert und anleitet (Bohnsack 2006). Daraus resultieren Gewohnheiten und Routinen, die sich auch in relativ stabilen, kollektiv erzeugten Kategorisierungen im Web dokumentieren (z.B. Golder & Huberman 2005). Kategorisierungen von Bildern auf Plattformen wie Flickr oder Instagram erlauben zudem, die Schlagworte den BildproduzentInnen zuzurechnen, da sie in der Regel die textliche Kategorisierung der Bilder vornehmen. Solche Kategorisierungen des Abgebildeten sollten sich daher im Besonderen dafür eignen, kultur- und webspezifische Wahrnehmungs-, Denk- und Bewertungsweisen aufzuzeigen. Einerseits könnten Bezeichnungen des Abgebildeten in Kombination mit anderen Schlagworten untersucht werden (Stichwort: Topic Modeling). Andererseits könnten die Prozesse der Kategorisierung nachverfolgt werden, um – für gewöhnlich unzugängliche – Prozesse von Neuverknüpfungen und die Entstehung neuartiger Sinnzusammenhänge zu beobachten. Eine entsprechende Visualisierung könnte ermöglichen, theoretisch postulierte Umkehrungen oder die „neuartige Fortsetzung von Eingelebtem“ (Hörning 2004: 33) sichtbar zu machen und empirisch in Feinanalyse zu untersuchen.

Die angesprochenen Kategorisierungsprozesse stellen besondere Anforderungen an automatisierte Visualisierungsverfahren, da sie nicht nur Sinnzusammenhänge in Form von Topics abbilden, sondern ebenfalls Verschiebungen im Orientierungswissen erfassen sollen. Das Poster formuliert kultursoziologische Ansprüche an entsprechende Softwarelösungen und diskutiert Potentiale und Grenzen von technischen Lösungen wie Topic Modeling zu unterschiedlichen Zeitpunkten, Cloudaliscious oder Alluvial Diagramme.

Referenzen

Bohnsack, R. (2006). Mannheims Wissenssoziologie als Methode. In: (Hg.): Neue Perspektiven der Wissenssoziologie (S. 271–291), herausgegeben von D. Tänzler, H. Knoblauch, & H.-G. Soeffner. Konstanz.

Geertz, C. (1987). *Dichte Beschreibung. Beiträge zum Verstehen kultureller Systeme.* Frankfurt a. M.

Golder, S. A., & Huberman, B. A. (2005). *The Structure of Collaborative Tagging Systems* (No. cs. DL/0508082). cs/0508082.

Hörning, Karl H. (2004). *Soziale Praxis zwischen Beharrung und Neuschöpfung.* In *Doing culture: neue Positionen zum Verhältnis von Kultur und sozialer Praxis* (S. 19-39), herausgegeben von K. Hörning, & J. Reuter. Bielefeld.

Manovich, L. (2011). *What is visualisation?* *Visual Studies*, 26(1), 36-49.

Mathes, A. (2004). *Folksonomies-cooperative classification and communication through shared metadata.* *Computer Mediated Communication*, 47(10), 1-13.

Mohr, J. W., & Bogdanov, P. (2013). *Topic models: What they are and why they matter.* *Poetics: Special Issue on Topic Models and the Cultural Sciences*, 41(6).

Neue Wege des Sammelns, Erfassens und Erforschens: Die Datenbank ‚Dialect Cultures‘.

Elisabeth Zehetner, Stefanie Edler

Institut für Germanistik, Karl-Franzens-Universität Graz, Heinrichstr. 26, 8010 Graz

Das Projekt ‚Dialect Cultures‘

Dialektliteratur erlebte im bairisch-österreichischen Raum bereits im 18. Jahrhundert eine erste Blüte, in der sie in ihrer inhaltlichen und funktionalen Bandbreite weit über jene idyllische, rückwärtsgewandte Heimatdichtung hinausging, mit der Mundartdichtung bis heute – auch in der Forschung – gerne identifiziert wird. Dialektkunst polarisierte wie keine andere Gattung das zeitgenössische Werturteil, wurde von Kaisern geliebt und von Kritikern verachtet, war in aller Munde und wurde von großen Meistern ebenso gepflegt wie von ungebildeten Laien.

Ihre Vielfalt ist jedoch kaum beachtet worden: Dialektliteratur führt in der Wissenschaft nach wie vor ein ungeliebtes Dasein, und die überlieferten Texte wurden bisher nicht systematisch dokumentiert oder kommentiert. Das Projekt ‚Dialect Cultures‘ will diese Lücke schließen und erforscht die verschiedenen ästhetischen und funktionalen Möglichkeiten der Dialektkunst im 17. und 18. Jahrhundert, indem bestehende Forschungsergebnisse und historische Quellen neu erschlossen und zusammengeführt werden. Die Materialgrundlage bildet dabei eine im Rahmen der ersten Projektphase erstellte umfassende Sammlung von historischen literarischen Texten und Notenmaterialien aus handschriftlichen oder gedruckten Quellen, welche nunmehr in einer Datenbank gebündelt, strukturiert und vernetzt vorliegt. Unter Berücksichtigung von Ansätzen aus unterschiedlichen Disziplinen soll Mundartverwendung auf dieser Basis als künstlerisches Phänomen vor 1800 in ihrer ganzen Bandbreite erfasst werden.

Die Datenbank

Kernstück des Projekts ist die Datenbank, in der im Rahmen der Projektarbeit seit 2010 literarische Texte des 17. und 18. Jahrhunderts gesammelt und kommentiert werden. Die Sammlung umfasst zurzeit ca. 1300 Werke aus den Bereichen Lyrik, Drama und Prosa in mehr als 2000 Varianten.

Werke lassen sich teilweise durch mehrere überlieferte Textzeugnisse belegen. Wenn diese untereinander leichte Abweichungen aufweisen, so sind für ein Werk mehrere Varianten zu verzeichnen, deren Differenzen in den jeweiligen Varianteneinträgen diskutiert werden. Die Varianteneinträge können mit entsprechend zugeordneten Autoren/Komponisten-, Quellen- sowie Literatureinträgen verlinkt werden. Darüber hinaus können den Varianten in gesonderten Dateien auch Digitalisate von Handschriften und Drucken sowie Transkriptionen zugeordnet werden. Für die Benutzer ist über die Variantenansicht auch der Zugriff auf diese Inhalte und somit z.B. ein Vergleich unterschiedlicher Varianten direkt am Originalmaterial möglich.

Auf einzigartige Weise verbindet die Datenbank so eine Datensammlung zu Texten historischer Dialektliteratur mit wissenschaftlicher Kommentierung und Edition. Diese bislang nur verstreut und in der Regel getrennt voneinander verfügbaren Informationen – in Bibliothekskatalogen und Überblicksdarstellungen einerseits, in Einzleditionen und Artikeln zu spezifi-

schen Themen andererseits – können so gesammelt und systematisch verknüpft werden. Alle Materialien und Informationen von der Quelle bis zur Forschungsliteratur sind damit auf einer gemeinsamen Plattform verfügbar.

Die Struktur der Datenbank ermöglicht auch das Aufdecken neuer Zusammenhänge, indem etwa verschiedene, bislang nicht bekannte Varianten verglichen werden oder thematische Schwerpunkte in der überlieferten Dialektliteratur und innerhalb einzelner Gattungen systematisch recherchiert werden können.

Die Datenbank erfüllt damit zwei wichtige Funktionen:

(1) Unterstützung der Forschung: Die online zugängliche Datenbank ermöglicht das Zusammenführen verschiedener Varianten und die Nachvollziehbarkeit von Quellen und Literatur, und dies insbesondere auch bei der Arbeit im Team. Der Datenbestand ist jederzeit von allen Beteiligten ausweitbar und für alle zeitgleich und übersichtlich nutzbar.

(2) Öffentlicher Zugang: Der Aufbau der Datenbank, der einen einfachen Zugriff über verschiedene Ebenen – Autoren, Werktitel und -incipits, Gattungen etc. – erlaubt, macht die Datensammlung über die Projektarbeit hinaus für ein breites Publikum nutzbar:

- Wissenschaftler aus verschiedenen Disziplinen wie Literaturwissenschaft, Sprachwissenschaft, Geschichte oder Musikwissenschaft, die mit ihren jeweils eigenen Fragestellungen an das Korpus herantreten können
- Studierende, die die Datenbank für Recherche und für das Kennenlernen von Transkriptions- und Editions-methoden nutzen können
- Interessierte außerhalb des Wissenschaftsbetriebs, für die die Ergebnisse wissenschaftlicher Forschung auf einfache Weise zugänglich werden.

Über eine Rückmeldefunktion können alle drei Gruppen nicht nur die Ergebnisse der Projektarbeit nutzen, sondern auch weiter ausbauen, indem neue Funde oder zusätzliche Kommentare zur Verfügung gestellt und – nach Überprüfung durch die Projektverantwortlichen – wieder in die Datenbank integriert werden können.

Die Datenbank gewährleistet also Offenheit und Austausch sowohl im Team als auch mit einem größeren Publikum und garantiert, dass die im Projektverlauf gesammelten Daten auch nach dem Ende der Projektlaufzeit gesichert und zugänglich bleiben.

Damit erweist sich das Modell der Datenbank mit ihrer Verknüpfung von verschiedenen Daten und Erkenntnissen als wegweisend über unser Projekt hinaus: Wissenschaft ist zunehmend durch eine wachsende Anzahl meist verhältnismäßig kurzfristiger Drittmittelprojekte gekennzeichnet, die häufig nur in geringem Ausmaß in die etablierten institutionellen Strukturen der Universität eingebunden sind. Gerade angesichts dessen scheint die längerfristige, umfassende Sicherung von Daten unerlässlich, um die Weiterverwendung der Ergebnisse und damit die Nachhaltigkeit des Erarbeiteten zu sichern.

neonion – Kollaboratives, semantisches Annotieren von Dokumenten als Mehrwert für das Forschen in den Geisteswissenschaften und der Informatik

Claudia Müller-Birn¹, Florian Schmaltz², Tina Klüwer¹, Juliane Stiller²

¹Freie Universität Berlin, Institut für Informatik, Human-Centered Computing

²Max-Planck-Institut fuer Wissenschaftsgeschichte, Forschungsprogramm Geschichte der Max-Planck-Gesellschaft

neonion ist eine Webanwendung, die es Benutzer_innen erlaubt, Wörter und Textteile in Dokumenten oder Dokumente selbst semantisch zu annotieren. Das Ziel bei der Softwareentwicklung ist es dabei insbesondere, den Prozess der Erstellung der semantischen Annotationen so intuitiv zu gestalten, dass die Komplexität des zugrundeliegenden Datenmodells vor den Nutzer_innen weitestgehend verborgen werden kann. Mit neonion sollen somit vor allem Wissenschaftler_innen angesprochen werden, die nicht mit semantischen Technologien wie RDF, Ontologien oder SPARQL vertraut sind, aber trotzdem von den Vorteilen dieser Technologien profitieren wollen. neonion ermöglicht es Dokumente gemeinschaftlich zu annotieren, d.h. als Mensch-Mensch oder Mensch-Maschine-Kollaboration. Annotationen können privat sein, in einer Gruppe gemeinsam erstellt oder sogar öffentlich sichtbar gemacht werden. So kann das bei der Annotation erstellte Wissen auch Teil des *Webs of Data* werden.

Derzeit wird neonion gemeinsam mit den Nutzer_innen aus dem Bereich der Geschichtswissenschaften entwickelt. In einem Pilotprojekt für das Forschungsprogramm „Geschichte der Max-Planck-Gesellschaft, 1948-2002“, welches am Max-Planck-Institut für Wissenschaftsgeschichte in Berlin angesiedelt ist, finden regelmäßige Treffen mit den potentiellen Nutzer_innen der Software statt.

Parallel zur Entwicklung der Software wird eine Studie durchgeführt, in der Wissenschaftler_innen interviewt werden und ihre Benutzung von neonion experimentell beobachtet wird. Erste Ergebnisse der Interviews zeigen, dass die dem Annotationsprozess zugrundeliegenden mentalen Modelle sehr unterschiedlich ausfallen, d.h. dass die Befragten ein individuelles Verständnis zum Konzept der Annotation besitzen. So wurde der verwendete Annotationsinhalt (z.B. Kategorie, Freitext, Tag) durch das Ziel der Auswertung (z.B. Klassifizierung, Übersetzung) beeinflusst. Die Befragten sehen semantische Annotationen dann als nützlich an, wenn Nutzer_innen ihre Annotationen gemeinschaftlich erstellen und verwenden sowie diese maschinell weiter verarbeiten können. In den Interviews hat sich gezeigt, dass Wissenschaftler_innen um die Nützlichkeit des semantischen Ansatzes wissen, aber sehr unsicher bei der eigentlichen Anwendung sind. Die fehlende Erfahrung in diesem Bereich führt letztlich wieder zur Verwendung von Werkzeugen wie beispielsweise Textverarbeitungsprogrammen oder PDF-Software, obwohl die Nachteile, beispielsweise bezüglich der Weiterverwendung der Annotationen, bekannt sind. Erste Usability-Studien mit neonion haben gezeigt, dass Nutzer_innen sich sehr gut in der Software zurechtfinden und Annotationen schnell und intuitiv durchführen können. Durch die enge Zusammenarbeit von Anwender_innen und Entwickler_innen können Möglichkeiten zur Verbesserung des Interaktionsdesigns frühzeitig im Entwicklungsprozess erkannt und adressiert werden.

Wie bereits dargelegt, ist ein Ziel von neonion die kollaborative Annotation von Dokumenten nicht nur für menschliche Interaktionen, sondern auch Mensch-Maschine-Interaktionen zu unterstützen. Dazu werden zwei Ebenen der manuellen semantischen Annotation unterschieden, die im Folgenden kurz anhand der Personenannotation erläutert werden:

(1) Auf der *Konzeptebene* annotiert der Nutzer_innen ausgewählte Wörter oder Textteile mit vorher festgelegten Begriffen (sog. Konzepte). Zum Beispiel sollen alle in historischen Dokumenten genannten Personen auf ihr gemeinsames Vorkommen (z.B. bezüglich Zeit und Ort) hin untersucht werden. Daher werden von den Nutzer_innen Namen, z.B. „Feodor Lynen“ mit dem Konzept „Person“ verbunden und damit im Dokument eine Annotation erzeugt. Die resultierende RDF-basierte Beschreibung der Instanz enthält die ausgewählten Namen vom Typ Person.

(2) Benutzer_innen können nicht nur die ausgewählten Namen auf ein Konzept in einer Ontologie beziehen, sondern diese Instanzen auch mit einer direkt identifizierbaren Ressource im Web verknüpfen. Wir bezeichnen dies als Annotation auf der *Referenzebene*. So könnte die lokal annotierte Person „Feodor Lynen“ mit dem Wikidata-Eintrag Q44597 zu Feodor Lynen referenziert werden.

Die Annotationen können nun um weiteres Wissen angereichert werden. Derzeit nutzt neonion Wikidata und dies erlaubt Nutzer_innen ebenfalls auf Daten aus der VIAF (Virtual International Authority File) oder der GND (Gemeinsame Normdatei) zuzugreifen. Indem also weitere Informationen aus der Linked Open Data Cloud einbezogen werden, kann auf Basis einer einfachen Annotation ein komplexes Wissensnetzwerk erzeugt werden. Solche Wissensnetzwerke können dann auch zur weiteren Analyse visualisiert werden.

Diese manuellen Annotationsebenen werden durch automatische Annotatoren, derzeit durch einen Named Entity Recognizer, erweitert. Während der manuellen Annotation von Dokumenten werden den Nutzer_innen Vorschläge des automatischen Annotators präsentiert. Die Nutzer_innen können diese annehmen, ablehnen oder editieren. Über diesen Interaktionsprozess werden zukünftig angebotene Empfehlungen schrittweise verbessert. Das entwickelte Mensch-Maschine-Interaktionskonzept wird derzeit evaluiert.

neonion soll als ein Beispiel für eine gelungene Zusammenarbeit zwischen den Geisteswissenschaften und der Informatik dienen, da die Forschung in beiden Fachdisziplinen mit diesem Projekt vorangetrieben werden kann.

Das Labeling System – ein freier Baukasten für kontrollierte Vokabulare

Michael Piotrowski
Florian Thiery
Kai-Christian Bruhn

10. November 2014

Maschinenlesbare Annotationen sind die Voraussetzung für die semantische Verarbeitung von Daten. Diese Aussage gilt unabhängig davon, ob es sich bei den Daten um natürlichsprachigen Text oder um strukturierte Datensätze in einer Datenbank handelt, und unabhängig davon, ob es um einfaches Sortieren und Filtern geht oder um komplexes automatisches Schließen. Kontrollierte Vokabulare (ob in einer einfachen Terminolgieleiste oder als Taxonomien, Thesauri oder Ontologien strukturiert) sind dabei unbedingt notwendig, um Annotationen maschinell verarbeitbar zu machen; ohne terminologische Kontrolle sind Annotationen für die maschinelle Verarbeitung kaum nützlicher als an den Rand eines Buches gekritzelte Notizen. Kontrollierte Vokabulare abstrahieren von natürlichsprachlichen Ambiguitäten und Konnotationen; sie sind daher entscheidend für die semantische Verarbeitung von Forschungsdaten. Um projektübergreifende Zusammenarbeit und den semantischen Austausch von Daten zu ermöglichen, müssen Vokabulare nicht nur kontrolliert, sondern auch formell oder informell standardisiert sein. Standardisierte kontrollierte Vokabulare ermöglichen den Austausch, die Kombination und die gemeinsame Analyse annotierter Daten aus verschiedenen Quellen sowie die Implementierung generischer Werkzeuge für die semantische Verarbeitung.

Erstellung und Wartung standardisierter kontrollierter Vokabulare sind jedoch zeitaufwändig und damit teuer. Zu den größten Herausforderungen zählen, dass alle beteiligten Parteien zu einem gemeinsamen Verständnis der Begriffe kommen, und dass die richtige Balance zwischen möglichst breiter Anwendbarkeit einerseits und möglichst präziser Analyse andererseits gefunden werden. Diese Ziele sind insbesondere in den Geisteswissenschaften schwierig zu erreichen: nicht nur sind die Forschungsfragen, die potentiell an einen gegebenen Datensatz gerichtet werden können, extrem weit gefächert, sondern die Kategorisierung der Daten ist häufig ein essenzieller Teil des Forschungsprozesses selbst. Es gibt daher einen eklatanten Mangel an standardisierten kontrollierten Vokabularen in den Geisteswissenschaften, der Digital-Humanities-Projekte letztlich dazu zwingt, eigene, projektspezifische Vokabulare zu definieren. Projektspezifische Vokabulare lösen können den internen Bedarf zwar kurzfristig befriedigen, sind aber nicht interoperabel und verhindern den zukünftigen Austausch und die Nachnutzung der annotierten Daten.

Unser Poster stellt einen neuen konzeptuellen Ansatz zur Lösung dieser Probleme vor und beschreibt die Implementierung dieses Ansatzes in einem Softwarewerkzeug, dem *Labeling System*.

Da es in der geisteswissenschaftlichen Forschung praktisch unmöglich ist, kontrollierte Vokabulare zu definieren, die alle denkbaren Anwendungen abdecken und generell akzeptiert sind, schlagen wir ein anderes Vorgehen vor. Bei unserem Ansatz definieren Projekte ihre eigenen Vokabulare, aber anstelle natürlichsprachlicher

Definitionen werden die Terme mit einem oder mehreren Konzepten in einem Referenzthesaurus verknüpft. Der projektspezifische Term dient also quasi als »Label« für eine Menge gemeinsamer Konzepte. Dieser Ansatz ermöglicht es Projekten Vokabulare entsprechend ihrer Bedürfnisse und unter Verwendung im jeweiligen Forschungsgebiet üblichen Bezeichnungen benutzen, während gleichzeitig die Interoperabilität mit anderen Projekten über den Referenzthesaurus gewährleistet ist.

Das Labeling System ist eine Webanwendung, die es Benutzern ermöglicht, SKOS-Vokabulare zu erstellen und auf einfache Weise deren Terme mit einem oder mehreren Konzepten in einem oder mehreren Referenzthesauri zu verknüpfen. Die Benutzeroberfläche ermöglicht die Visualisierung der definierten Vokabulare in einer hierarchischen Baumstruktur und ermöglicht den Zugriff auf Vokabulare über eine SPARQL-Schnittstelle. Das Labeling System basiert auf ausgereiften Open-Source-Komponenten und ist selbst ebenfalls frei verfügbar.

Learning cuneiform the modern way

Timo Homburg¹, Christian Chiarcos¹, Thomas Richter², Dirk Wicke²

¹ Institute for Computer Science ² Institute for Archaeology
Goethe University, Frankfurt, Germany
`timo.homburg@stud.uni-frankfurt.de`,
`{chiarcos|thomas.richter|wicke}@em.uni-frankfurt.de`

Keywords: Assyriology, cuneiform, input method engines (IME), flash card learning

With our poster and the accompanying demo, we present current progress on the information-technological support for scholars and students of cuneiform. For a period of about 3000 years, cuneiform was the dominant writing system of the Ancient Near East, with a rich literary tradition in several languages, and an extensive amount of texts preserved in tens of thousands of clay tablets.

Despite this wealth of data and a strong academic tradition in their analysis, the numerous specific challenges of cuneiform have only partially been addressed so far. Here, we propose adapting input method engines (IMEs) and learning strategies commonly used for Asian languages according to the needs of Assyriology.

Typing cuneiform Cuneiform writing for Akkadian, Sumerian and Hittite is ideosyllabic, i.e., combining syllabic and ideographic elements, often for the same sign. Up until this date there is no free and convenient way of typing Unicode cuneiform characters other than utilizing the Unicode code tables directly, i.e., to copy and paste from online dictionaries. Not all online dictionaries, however, use Unicode symbols, some use legacy fonts, some represent signs by images, and some do not provide a cuneiform representation at all.

To accommodate this deficit, we developed an input method which is based on the transliteration concept of Chinese Pinyin, the most common way of typing non-alphabetical languages on a computer. To achieve an equivalent input for the aforementioned languages we utilized a given char transliteration to cuneiform table¹ to create transliteration to cuneiform mappings of Akkadian, Sumerian and Hittite CDLI² corpora respectively. Organized as a tree, thus minimizing latency, word and char-based input method engines were created for Java (JIMF,

¹ <http://www.acoli.informatik.uni-frankfurt.de/resources/cuneiform/signs-final.xml>

² <http://cdli.ucla.edu/>

Fig. 1.a)³, JQuery(Fig. 2)⁴, SCIM⁵ and Ibus (Fig. 3)⁶, thereby covering the most important input method engines on Linux, Web and Java environments.

Learning cuneiform Because of limited technological support, digital resources in assyriology often focus on transliteration or transcription as means of representation whereas students are required to acquire the necessary knowledge on cuneiform characters on their own. Clearly, none of those practices are satisfying or easily adaptable for text processing and therefore not useful for computer-aided teaching methods. For conveniently typing and learning cuneiform characters, words and phrases for the Akkadian, Sumerian and Hittite language, we present an adaptation of Anki, a common tool for flash card learning (Fig. 1.b). We utilized the existing character/word table to create flash card sets consisting of more than 50000 words for the Anki and AnkiDroid⁷ flash card learning program. Subsequent extensions may exploit existing corpora, e.g., the Open Richly Annotated Cuneiform Corpus (ORACC),⁸ to create flash cards for words and phrases. Anki schedules learning content according to a spaced repetition learning method having proven its positive learning effect over a longer period of time to maximize learning success. Because of its usability in both mobile and desktop environments and its ability to share flashcards online Anki suits not only the students but also simplifies sharing lecture specified flash card sets for the lecturers.

In our presentation, we demonstrate how an input method engine can act as a suitable tool for solving the mentioned input and compatibility problems while at the same time being useful for education and language learning purposes. With the tools described above, teachers can now easily create their own flash cards according to the pace and content of their lectures. Students may enjoy a convenient and scientifically proven way of learning cuneiform vocabulary, as well as a way to prove their learning by utilizing the input method engine to create their own cuneiform texts. In conclusion, a notable improvement in writing and in learning the concerned languages has been realized and is in general perceived well.

Both tools and the accompanying sign/word table have been created in the context of on-going experiments on word segmentation and transliteration in cuneiform languages. In this regard, we are thus primarily working on *processing cuneiform*. In addition to demonstrating input methods and learning tools, we will include early results with respect to these aspects in demo and presentation, as well.

³ <http://docs.oracle.com/javase/7/docs/technotes/guides/imf/overview.html>

⁴ Sourcecode: <https://github.com/situx/webime>

Livedemo: <http://www.web-ime.de.vu>

⁵ <http://sourceforge.net/projects/scim/>

⁶ <https://code.google.com/p/ibus/>

⁷ <http://ankisrs.net>

⁸ <http://http://oracc.museum.upenn.edu>

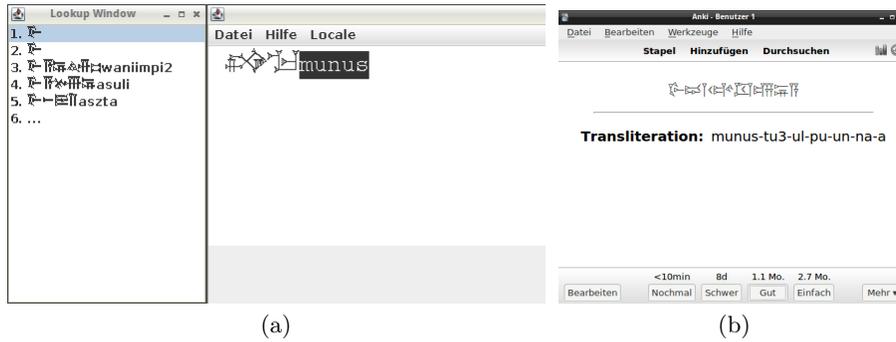


Fig. 1. Learning Cuneiform: (a) Java Input Method Framework based IME for Swing based applications and (b) Anki Flash Cards

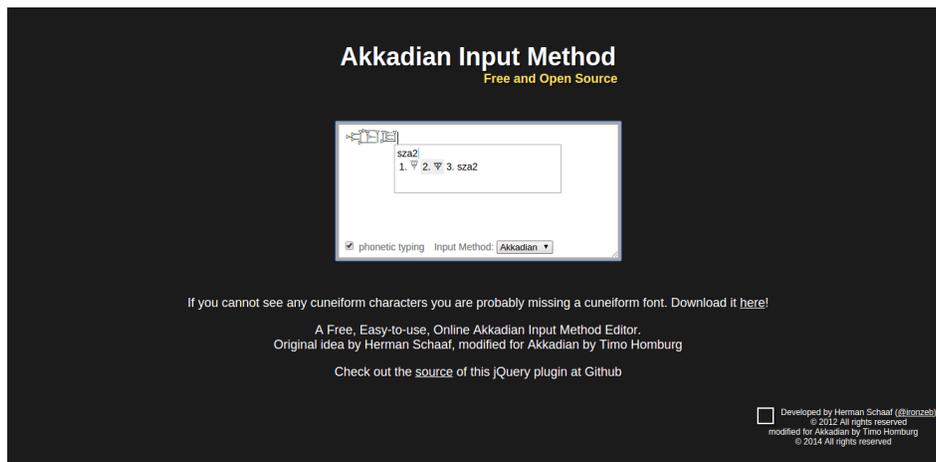


Fig. 2. JQuery based Akkadian Input Method Engine testable on <http://www.webime.de.vu>

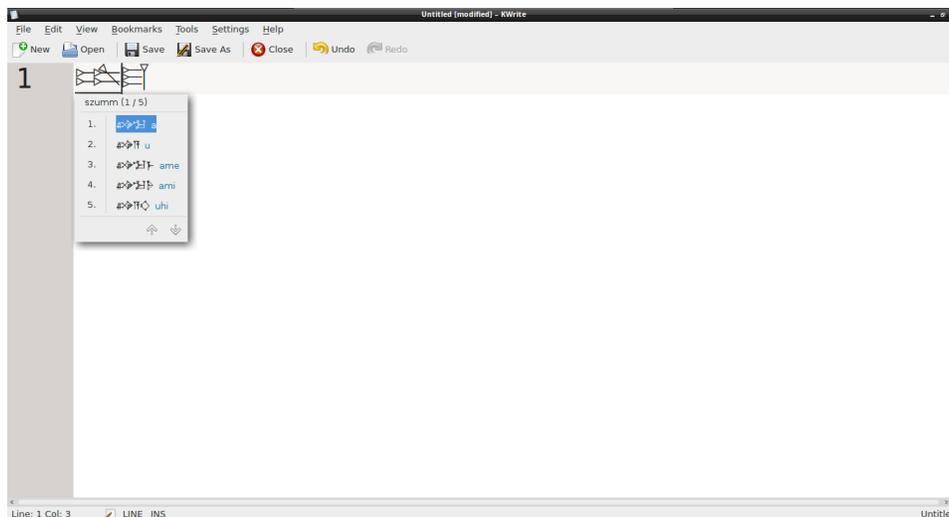


Fig. 3. System-wide Input Method Engine using Ibus for Linux with SCIM giving a similar output

Modellierung eines maschinell lesbaren Lexikons für das Korpus der altäthiopischen Literatur

Alessandro Bausi, Andreas Ellwardt, Cristina Vertan
Universität Hamburg

1. Einführung

Die Entwicklung und ständige Erweiterung des Unicode-Kodierungssystems Unicode¹ sowie der Mark-up-Sprachen XML², TEI³ haben in den letzten Jahren u.a. die digitale textuelle Repräsentation von historischen Dokumenten, die mit unterschiedlichen Alphabeten geschrieben wurden, ermöglicht.

Diese textuelle Repräsentation eröffnet wiederum, im Kontrast zur reinen Speicherung von Bild-Digitalisaten, die Möglichkeit, computergestützte linguistische sowie philologische Untersuchungen auf großen Textmengen durchzuführen. Durch solche Methoden lässt sich beispielsweise eine diachrone Analyse der Sprache gleichzeitig auf mehreren Ebenen (morphologisch, syntaktisch, semantisch) realisieren, vorausgesetzt, die elektronischen Ressourcen wie Lexika oder annotierte Korpora sowie die sprachtechnologischen Prozesse (morphologische Analytiker, Wortart-Tagger, Parser) sind vorhanden.

Während die sprachtechnologischen Ressourcen und Werkzeuge für moderne Sprachen sehr weit entwickelt sind, gelten viele historische Sprachen als stark „under-resourced“. Laut Krauwer (2003) gibt es ein minimales Set von Ressourcen, die für eine computergestützte Sprachanalyse unabdingbar sind. Dessen Weiterentwicklung stellt die Wissenschaft vor neue Forschungsprobleme, da sich häufig Modelle, die für moderne Sprachen entwickelt wurden, nicht 1:1 auf historische Sprachen übertragen lassen (VertanEtAl.2014)

In diesem Beitrag werden wir die Modellierung und Entwicklung von sprachtechnologischen Ressourcen für das Altäthiopische (Ge‘ez) erläutern. Die Besonderheiten des Ge‘ez (s. Sektion 2), bedingen die Entwicklung von neuen Modellen, z.B. im Bereich der Lexika. In Sektion 3 werden wir exemplarisch die Entwicklung eines Lexikon-Modells für Ge‘ez darstellen, während wir in Sektion 4 die Einbindung des Lexikons in einer Architektur für die diachrone Analyse des Ge‘ez diskutieren werden.

2. Kurze Darstellung des Altäthiopischen (Ge‘ez)

Das südsemitische Ge‘ez ist die Sprache des Königreichs Aksum in der heutigen nordäthiopischen Provinz Tigray, von wo aus die im 4. Jahrhundert beginnende Christianisierung Äthiopiens ihren Anfang nahm. Die in der Folge entstehende reiche Literatur ist in großem Umfang geprägt von Übersetzungen aus dem Griechischen und später, ab dem 13. Jahrhundert, aus dem Arabischen, was durch grammatische Interferenzphänomene reflektiert wird. Während seine Verdrängung als gesprochene Sprache bereits im 9./10. Jahrhundert beginnt, bleibt es als Schriftsprache sehr viel länger erhalten und ist bis in die Gegenwart hinein Liturgiesprache des äthiopischen und eriträischen Klerus.

Das Altäthiopische hat aus einer südsemitischen Schrift ein eigenes Silbenalphabet entwickelt, das bis heute in mehreren modernen Sprachen Äthiopiens und Eritreas Verwendung findet. Innerhalb der semitischen Sprachen fällt es durch die verwendete Rechtsläufigkeit auf, außerdem werden die Vokale vollständig geschrieben. Beides unterscheidet das Ge‘ez von den ihm nächst verwandten Sprachen Altsüdarabisch, Arabisch, Hebräisch und Syro-Aramäisch. Des weiteren sind Grapheme, die ursprünglich distinkten Phonemen zugeordnet waren, schon früh in identischer phonetischer Realisierung zusammengefallen, was sich konkret bereits in den ältesten überlieferten Handschriftzeugnissen (aber noch nicht in den aksumitischen Inschriften) niederschlägt, wo eine beliebige Austauschbarkeit der Laryngale und Sibilanten jeweils untereinander zu konstatieren ist.

Mit den genannten eng verwandten semitischen Sprachen teilt das Altäthiopische die nichtkonkatenative Morphologie. Hierbei muss das einzelne Lexem als Kombination von zwei Elementen beschrieben werden, nämlich der Wurzel und dem Schema: Die konsonantische Wurzel gibt veränderliche Positionen zwischen

¹ <http://www.unicode.org/>

² <http://www.w3.org/XML/>

³ <http://www.tei-c.org/index.xml>

ihren, zumeist drei, Wurzelkonsonanten vor, die durch die Vokale des Schemas aufgefüllt werden, häufig, jedoch nicht zwingend, ergänzt um (vokalische oder konsonantische) Affixe.

3. Arbeitsschritte zu einer computergestützten Analyse des Altäthiopischen

Wie bereits in Sektion 2 erwähnt, sind Ge'ez-Dokumente für die gesamte Geschichte des christlichen Orients extrem wertvoll. Manche Überlieferungen von alten griechischen Texten sind in der Originalsprache verloren und nur im Altäthiopischen erhalten. In der Zeit digitaler Bibliotheken erscheint also die Entwicklung von computergestützten Tools für die Ge'ez-Sprache umso dringender. Das primäre Ziel des Projekts TraCES⁴ ist die Entwicklung eines digitalen Korpus der Ge'ez-Sprache, zusammen mit Annotationen auf morphologischer, syntaktischer und semantischer Ebene. Dieses annotierte Korpus soll einerseits eine diachrone Analyse des Altäthiopischen ermöglichen, andererseits soll es selbst als Ressource für weitere computergestützte Prozesse dienen. Langfristig soll eine vergleichende digitale Analyse von altäthiopischen und griechischen (z.B. die in der digitalen PERSEUS Sammlung⁵ verfügbaren) oder arabischen sowie anderen christlich-orientalischen Dokumenten möglich sein.

Mit Ausnahme von einigen wenigen Texten gibt es zur Zeit keine verfügbare elektronische Ressource für das Altäthiopische. Daher haben wir uns als erstes der Entwicklung eines maschinell lesbaren Lexikons des Ge'ez gewidmet. Dessen Modellierung wird in der nächsten Sektion erklärt.

4. Ein Lexikon-Modell für Ge'ez

Die in Sektion 2 erwähnte Austauschbarkeit der Laryngalen und Sibilanten untereinander stellt uns vor eine erste Modellierungsanforderung. Für einen Lexikon-Eintrag muss nicht nur die Grundform, sondern es müssen auch alle möglichen graphischen Varianten gespeichert werden, wobei wohlgermerkt diese graphische Variationen auch in einigen Fällen als selbständige Lexikon-Einträge mit ganz anderer Bedeutung existieren können.

Das Lexikonmodell muss daher eine starke Modularisierung und Verlinkung zwischen den einzelnen Modulen unterstützen. Wir haben uns für das Lemon-Modell (McCraeEtAl.2012) entschieden. Unserer Kenntnis nach, ist dies der erste Versuch, eine semitische Sprache mit dem Lemon-Modell zu beschreiben. Die Grundkomponenten eines Lemon-Lexikon-Modells für Ge'ez wurden wie folgt angepasst.

Die Zitierform eines Wortes in klassischen Lexika semitischer Sprachen ist in der Regel eine verbale Repräsentation der Wurzel in der 3. Person Perfekt Singular maskulin. Diese Form wird in unserem Lemon-Modell als „Lexical Entry“ gespeichert.

Ein „Lexical Entry“ ist mit den folgenden weiteren Modulen verknüpft:

- Das Lexical Form-Modul beinhaltet alle möglichen graphischen Varianten des Lemmas. Jede graphische Variante wird zusammen mit ihrer Transkription gespeichert.
- Das Morphologie-Modul beinhaltet eine Subkomponente für den lexikalischen Eintrag, die das Paradigma, Ausnahmen der morphologischen Realisierung (z.B. Sonderformen im Imperfekt oder Plural) sowie die jeweiligen anderen morphologischen Kategorien für das Lemma umfasst. Das Semantik-Modul setzt sich aus einer Übersetzungs-, einer Korpusevidenz- und einer semantische-Merkmale-Komponente zusammen. Unter Korpusevidenz verstehen wir Beispiele aus Korpora für dieses Lemma oder eine seiner morphologischen Realisierungen. Die Übersetzungen sind unterteilt in eine Übersetzung ins Englische und semantische Äquivalente in anderen Sprachen wie (falls vorhanden) Arabisch, Hebräisch, Syrisch, Koptisch, Griechisch oder sogar Sanskrit.
- Das Syntax-Modul beinhaltet syntaktische Funktion des Lemmas, zusammen mit Beispielen von syntaktischen Bäumen. Dieses Modul wird in einer späteren Projektphase entwickelt.

⁴ European Union Seventh Framework Programme IDEAS (FP7/2007-2013), European Research Council, grant agreement no. 338756, project “TraCES – From Translation to Creation: Changes in Ethiopic Style and Lexicon from Late Antiquity to the Middle Ages”, <http://www1.uni-hamburg.de/ethiostudies/traces.html>

⁵ <http://www.perseus.tufts.edu/hopper/>

4.1. Wurzel-Modellierung

Da die Wurzel eine zentrale Stellung in der semitischen Morphologie hat, haben wir als ersten Schritt ein Wurzel-Sublexikon erstellt. Dieses entspricht dem Wurzel-Submodul im morphologischen Modul.

Die Erstellung des Wurzel-Lexikons wurde vollständig automatisiert. Aus einer digitalen Version des trotz seiner Abfassung im Jahre 1865 unverändert als Standardwerk geltenden „Lexicon linguae aethiopicae“ von August Dillmann (Dillmann1865) (im Unicode-Format) wurden zirka 4000 Wurzel-Einträge mit Hilfe von String-basierten Regeln extrahiert.

Für jede Wurzel wurden:

- die vollständige Transkription
- die auf das konsonantische Gerüst zurückgeführte Transkription
- das konsonantischen Wortbildungsschema
- alle graphischen Varianten zusammen mit deren Transkriptionen

durch regel-basierte Verfahren extrahiert. Die Automatisierung ermöglicht zum ersten Mal die Sammlung aller graphischen Varianten für alle 4000 Wurzeln (wobei hervorgehoben werden muss, dass manche Wurzeln bis zu 50 graphische Varianten haben).

Jede Wurzel wird automatisch mit ihren Homophonen (Einträge mit identischer graphischer Form, aber unterschiedlicher Bedeutung) verknüpft. Erfasst werden durch automatische Prozesse auch alle Lexikoneinträge von graphischen Varianten (falls vorhanden).

Das Wurzel-Lexikon wird im XML-Format gespeichert. Dafür wurde ein eigenes XML-Schema entworfen. Eine Java-basierte graphische Oberfläche wurde implementiert. Diese Oberfläche ermöglicht nicht nur die Visualisierung von den Einzeleinträgen und die Navigation durch das Wurzel-Lexikon, sondern auch manuelle Korrekturen, das Löschen oder das Einfügen von neuen Einträgen.

Nach Korrekturen wird das Wurzel-Lexikon:

- als eine „Authority List“ für das Ge‘ez-Lexikon und
- als Generierungsquelle für Lexikoneinträge

benutzt.

5. Zusammenfassung und weitere Arbeit

In diesem Beitrag haben wir die Modelle für ein Wurzel- und ein Lemma-Lexikon für die Ge‘ez-Sprache erklärt. Die Wurzel und Lemma-Akquisition werden weitgehend durch computergestützte Prozesse realisiert. Die erstellte Software wird bei der Präsentation des Beitrags vorgeführt.

Das Projekt TraCES wurde im März 2014 begonnen und hat eine Laufzeit von fünf Jahren. Die Erstellung des Lexikons der Ge‘ez Sprache ist zurzeit die zentrale Arbeit im Projekt, wobei derzeit die Erstellung von Generierungsparadigmen im Vordergrund steht. Mit deren Hilfe werden durch Computerverfahren Lexikoneinträge generiert.

Ein erster Test hat mehr als 13 000 Einträge generiert. Dies zeigt, dass die Automatisierung eine erhebliche Zeitersparnis für die Lexikon-Akquisition ermöglicht.

Literatur

(Dillmann1865) Dillmann, August, *Lexicon linguae Aethiopicæ cum indice Latino*, Lipsiae 1865.

(Krauer2003) Krauer, Steven, „*The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources*“, <http://www.elsnet.org/dox/krauer-specem2003.pdf> (09.11.2014)

(McCraeEtAl2012). McCrae, John und Aguado-de-Cea, Guadalupe und Buitelaar, Paul und, Cimiano, Philipp und Declerck, Thierry und Gómez Pérez, Asunción und Gracia, Jorge und Hollink, Laura und Montiel-Ponsoda, Elena und Spohr, Dennis und Wunner, Tobias, *The Lemon Cookbook*, <http://lemon-model.net/lemon-cookbook.pdf> (09.11.2014)

(VertanET.AL.2014) Vertan, Cristina und Zervanou, Kalliopi und van den Bosch, Antal und Sporeleder, Caroline (Hrsg.), *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Association for Computational Linguistics, Götheborg, Sweden, 2014, <http://www.aclweb.org/anthology/W14-06> (09.11.2014)

Titel: OpenSource-Bibliotheken und -Tools des SeNeReKo-Projekts

Autoren: Jürgen Knauth, Frederik Elwert

Abstract

Ziel des SeNeReKo-Projektes ist es, durch Techniken der **Semantisch-Sozialen Netzwerkanalyse** einen Einblick in Teile von Textkorpora zu erhalten. Damit wird eine Form des „Distant Readings“ realisiert, im konkreten Fall zur Erforschung von **Religionskontakten** in altägyptischen Texten und dem Pali-Kannon. (= SeNeReKo)

Im Kontext des Projekts sind verschiedene Programmierbibliotheken und Werkzeuge entwickelt worden, um den Anforderungen des Projekts gerecht zu werden. Wesentliche Komponenten sind jedoch nicht projektspezifisch, sondern als allgemein verwendbare OpenSource-Komponenten geplant und umgesetzt worden: Wiederverwertbarkeit war von Vorneherein eines der Entwicklungsziele. Da Teilaufgaben von DH-Projekten durchaus öfters ähnlich gelagert sind, ist davon auszugehen, dass die so entstandenen Komponenten und Tools von anderen Wissenschaftlern entweder direkt oder nach geringer Adaption für andere Projekte genutzt werden können: Ziel des vorliegenden Posters ist es daher über genau diese Komponenten und Werkzeuge zu informieren. Da unsere Werkzeuge gerade deswegen entstanden sind, weil bislang noch nichts Vergleichbares zur Verfügung stand um die von uns angetroffenen Probleme effizient zu lösen, hoffen wir so durch unsere Software einen Beitrag für die Wissenschafts-Community zu leisten und so andere Wissenschaftler in ihrer zukünftigen Arbeit unterstützen zu können.

Konkret wurde in SeNeReKo ein Werkzeug zum Tagging von Texten entwickelt. Eine Besonderheit dieses Werkzeugs ist neben der Eigenheiten zur Auflösung von Pali-Sandhis seine besonders gut optimierte Usability: Unser Anliegen war hier, dass möglichst wenige Klicks erforderlich werden, um manuelle Tagging-Aufgaben durchzuführen. Das Fehlen von Tools mit vergleichbar Usability war Motivation der Entwicklung dieses Werkzeugs. Dieses client-server-basierte Standalone-Tool leistete einen wertvollen Beitrag in SeNeReKo für die Erstellung eines Gold-Standards im Pali, um weitere computerlinguistische Arbeitsschritte zu ermöglichen. Das Tool selbst kann jederzeit an die Verwendung für andere Sprachen angepasst werden.

Ferner stellen wir einen NoSQL-basierten Server zur Verwaltung von Wörterbuchdaten vor. Dieser ist als Komponente in einer klassischen Client-Server-Umgebung konzipiert und wird von den gleich nachfolgend erwähnten Werkzeugen verwendet: Ein Tool zur maschinellen Verarbeitung dieser Wörterbucheinträge, sowie einem Tool zur Visualisierung einzelner Datensätze. Der Server verwaltet dabei alle Wörterbucheinträge zentral und erlaubt dank seiner Bulk-Requests das effiziente Durchforsten der Daten auch bei größeren Datenmengen.

Ein Transformationswerkzeug, welches an den Server andockt, ist als IDE (Integrated Development Environment) konzipiert: Es erlaubt die Eingabe von C#-Programmcode-Fragmenten zur Datenverarbeitung. Diese Fragmente werden kompiliert; dann können sämtliche Wörterbucheinträge mit diesem Kompilat verarbeitet werden, um z.B. Muster zu erkennen und darauf basierend einzelne Wörterbucheinträge mit erkannten Informationen anzureichern. Eine Preview-Funktion gibt genauen Einblick darüber, auf welche Einträge sich die aktuell eingegebene Verarbeitungslogik erstreckt. Dadurch entsteht Transparenz: Erst der Einblick in die konkreten Änderungen über alle Datensätze hinweg erlaubt eine effiziente weil fehlerfreie Überarbeitung von Wörterbucheinträgen.

Ebenfalls an den Wörterbuchserver angegliedert ist ein Werkzeug zur Suche und Darstellung einzelner Wörterbuchartikel. Per serverseitig gespeicherter Konfiguration kann festgelegt werden, welche Controls in der GUI angezeigt werden sollen, und mit welchen Datenfeldern der einzelnen Artikel diese verbunden sein sollen: So kann eine Anpassung an Wörterbuchdaten beliebiger Struktur mit wenigen Handgriffen erfolgen. Die graphische Oberfläche erlaubt es, die manuelle Überarbeitungen auch größerer Artikelmengen auf einfachem Weg zu realisieren.

Ein anderes SeNeReKo-Tool unterstützt die Transformation beliebiger XML-Daten nach TEI: Über eine IDE-ähnliche Oberfläche können Umwandlungsregeln in Form eines Skripts eingegeben werden. Diese sind so gestaltet, dass sie fast schon natürlichsprachlich und somit leicht verständlich sind. Eine Erweiterbarkeit durch eigene Regeln ist jederzeit möglich: So können auch projektspezifische und über klassische X-Technologien möglicherweise nur schwer realisierbare Verarbeitungsprozesse durch ein Kommando repräsentiert werden (wie u.a. zur Verarbeitung im Pali in SeNeReKo). Angewandt auf eingelesene XML Datei(en) kann so die Aufbereitung von Daten erleichtert werden.

Des Weiteren stellen wir auf unserem Poster eine Reihe Tools vor, die dazu verwendet werden können, um aus TEI- bzw. TCF-Daten Netzwerke zu erzeugen. Sie stellen die Basis für den Kern des Projekts – die Erzeugung von Netzwerken dar. Da hier Standard-Graph-Datenformate für die Ausgabe verwendet werden, ist es möglich, diese Netzwerke dann anschließend mit verfügbaren Standard-Tools zu visualisieren.

Diese oben genannten Werkzeuge (bzw. Bibliotheken) sind frei nutzbar und helfen, gerade den teilweise sehr schwierigen Prozess der Datenaufbereitung zu adressieren. Wir würden uns freuen, wenn diese Komponenten nicht nur uns in SeNeReKo, sondern zukünftig auch anderen Wissenschaftlern helfen, damit so genau die Brücke geschlagen werden kann, die im Zentrum der Tätigkeit von uns Wissenschaftlern liegt: Die Brücke von Daten zu Erkenntnissen.

Annotationen für die automatisierte Verarbeitung von Märchen

Thierry Declerck, Universität des Saarlandes

(word count: 741)

In diesem Poster- und Demobeitrag fassen wir ältere und aktuelle Arbeiten zur Entwicklung eines Annotationsschemas für Märchen zusammen, das auch die Einbettung von Märchentexten in automatisierten Verarbeitungszenarien erlaubt. Eine Entwicklung unserer Arbeit in diesem Bereich führte zur automatischen Erkennung von Charakteren in Märchen, deren Rolle in Dialogen und deren Emotionen, die als Grundlage eines TextToSpeech Szenarios dient, das Märchentexte „vorliest“.

Dieses Ergebnis basiert auf einer Zusammenarbeit mit Studenten der Computerlinguistik an der Universität des Saarlandes, die in den letzten Jahren in Form von Bachelor- oder Masterarbeiten, oder auch in Form eines Softwareprojekts erfolgten.

Angefangen hat es mit der Masterarbeit von Antonia Scheidel zur Annotation von Märchen mit Proppschen¹ Funktionen. Antonia Scheidel entwickelte ein neues Annotationsschemas, nach dem Märchen nach Texteigenschaften, temporalen Strukturen, Charakteren, Dialogen, und Proppschen Funktionen abfragen kann (s. [1]). Ein Annotationsschema ist insofern wichtig, als dadurch automatisierte Systeme ein Ziel haben, in das sie ihre Ergebnisse abbilden können. Wenn dazu auch Märchen mit dem Annotationsschema manuell annotiert werden, können die Ergebnisse der automatischen Verarbeitungen mit den menschlichen Annotationen verglichen werden.

Darauf aufbauend hat Nikolina Koleva an einem automatisierten System gearbeitet, das in Märchentext (sie hat mit 2 Beispielen gearbeitet; „The Magic Swan Geese“, eine englische Version eines russischen Märchens, und „Väterchen Frost“, eine deutsche Version eines russischen Märchens). Sie hat ein Programm geschrieben, das den Text nach linguistischen Kriterien analysiert, mit dem Ziel, die darin vorkommenden Charaktere zu erkennen, und in eine Datenbank zu speichern. Diese Datenbank ist von der Sorte „Ontologie“: darin können logische Operationen durchgeführt werden. Als Hintergrund fungiert eine formale Beschreibung dessen, was in den genannten Märchen vorkommen kann, inklusive eine Ontologie über Familienverhältnissen. So kann das System erkennen, dass im Text „die Tochter“ die gleiche Person wie die „Schwester“ ist, wenn der Kontext dies suggeriert. Erkannte Charaktere im Märchen werden somit mit allgemeineren Kategorien

¹ Auszug aus Wikipedia: „Propp gilt als Begründer der morphologischen oder strukturalistischen Folkloristik. Zwischen 1914 und 1918 studierte er russische und deutsche Philologie. Danach unterrichtete er die deutsche Sprache an verschiedenen Hochschulen in Leningrad. Von 1938 bis 1969 war er Professor für Germanistik, russische Literatur und Folklore an der Staatlichen Universität Leningrad.

1928 erschien sein bahnbrechendes Werk *Morphologie des Märchens*. Das Buch wurde 1958 in den USA in englischer Sprache veröffentlicht, was Propp weltweite Anerkennung verschaffte. 1946 erschien das Buch *Die historischen Wurzeln des Zaubermärchens*.”

(http://www.wikiwand.com/de/Wladimir_Jakowlewitsch_Propp. Zugriff am 2014.11.10)

semantisch annotiert. Und wir wissen dann in welchen Kontexten (oder Situationen) die Tochter (zum Beispiel) involviert ist (s. hierzu [2]).

Schließlich eine Gruppe von Studenten (Christian Eisenreich, Jana Ott, Tonio Süßdorf und Christian Wilms) im Rahmen eines Softwareprojekts an Erweiterungen der oben genannten Arbeiten gearbeitet. Sie haben zum einen das Annotationsschema erweitert, mit detaillierteren Dialogbeschreibungen, und mit der Kodierung von Emotionen. Die Ontologie wurde auch erweitert, und sie inkludiert jetzt auch eine Beschreibung von Dialogen (Fragen, Antworten, Monologe, etc.), inklusive der Kodierungen der Teilnehmern und der Dialogwechseln. Auch 6 Basisemotionen (Angst, Trauer, Freude, etc) sind in der Ontologie kodiert.

Eine Haupterweiterung der vergangenen Arbeiten besteht darin, dass auch synthetische Stimmen eine Rolle spielen. Ist einmal ein Charakter erkannt worden, zum Beispiel die Prinzessin (im Märchen „Froschkönig“), werden zusätzliche Merkmale kodiert (zum Bsp. Alter, usw.). Dann wird automatisch eine vorher definierte synthetische Stimme zum Charakter addiert. Wenn dann der Text von dem System analysiert wird, kann die Geschichte von den Stimmen „erzählt“ werden. Wenn kein Charakter in einer Dialogsituation vorkommt, dann wird angenommen, dass der Erzähler/die Erzählerin „daran“ ist. Eine Demo kann hier gehört werden:

https://bytebucket.org/ceisen/apftml2repo/raw/763c5eb533f09997e757ec61652310c742238384/example%20output/audio_output.mp3

Im Anhang sind 2 Screenshots, die (für den ersten Teil der Audiodatei) zeigen wie das System den Text bearbeitet und kodiert, so dass die Sprachausgabe (s. Link oben) erzeugt werden kann. Unser Poster/Demo zeigt die Korrelation zwischen die Annotationen, die zum größten Teil automatisch generiert worden sind, und den verschiedenen Stufen der Verarbeitung bis hin zur Sprachausgabe.

Referenzen

[1] Thierry Declerck, Antonia Scheidel, Piroska Lendvai. **Proprian Content Descriptors in an Integrated Annotation Schema for Fairy Tales.** *Language Technology for Cultural Heritage. Selected Papers from the LaTeCH Workshop Series, Theory and Applications of Natural Language Processing, Pages 155-169, Springer, Heidelberg, 2011*

[2] Nikolina Koleva, Thierry Declerck, Hans-Ulrich Krieger. **An Ontology-Based Iterative Text Processing Strategy for Detecting and Recognizing Characters in Folktales** in: Jan Christoph Meister (ed.): *Digital Humanities 2012 Conference Abstracts, Pages 467-470, Hamburg.*

[3] Christian Eisenreich, Jana Ott, Tonio Süßdorf, Christian Willms, Thierry Declerck. **From Tale to Speech: Ontology-based Emotion and Dialogue Annotation of Fairy Tales with a TTS Output** *Proceedings of ISWC 2014, Riva del Garda, Italy, Springer.*

Anhang

```
ca. Command Prompt - run_ja.bat
...finished
building and writing xml...
...finished
populating ontology...
...finished
generating ITS script from ontology...
...finished
computing and playing audio...
-----
narrator added
      ID: -1
-----
[-1] in olden times, when wishing still did some good, there lived a king whose
      daughters were all beautiful, but the youngest was so beautiful that the sun it
      self, who, indeed, has seen so much, marveled every time it shone upon her face.
[-1] in the vicinity of the king's castle there was a large, dark forest, and in
      this forest, beneath an old linden tree, there was a well.
[-1] in the heat of the day the princess would go out into the forest and sit on
      the edge of the cool well.
[-1] to pass the time she would take a golden ball, throw it into the air, and
      then catch it.
[-1] it was her favorite plaything.
[-1] now one day it happened that the princess's golden ball did not fall into
      her hands, that she held up high, but instead it fell to the ground and rolled r
      ight into the water.
[-1] the princess followed it with her eyes, but the ball disappeared, and the
      well was so deep that she could not see its bottom.
[-1] <sad>then she began to cry.
[-1] <sad>she cried louder and louder, and she could not console herself.
[-1] <sad>as she was thus lamenting, someone called out to her,
-----
sender added
      ID: 2
      Attributes: [Animal, Character, Sender, Receiver, Frog, Physical]
      Voice: EN_FROGLIKE
-----
[2] <sad>what is the matter with you, princess? your crying would turn a stone
      to pity.
[-1] she looked around to see where the voice was coming from and saw a frog, w
      ho had stuck his thick, ugly head out of the water.
-----
sender added
      ID: 1
      Attributes: [Human, BiolDaughter, Character, Daughter, Sender, Receiver,
      Girl, Physical]
      Voice: EN_TEENAGE_FEMALE
-----
[1] oh, it's you, old water-splasher.
[-1] she said .
[1] <sad>i am crying because my golden ball has fallen into the well.
[2] <sad>be still and stop crying,
[-1] answered the frog .
[2] i can help you, but what will you give me if i bring back your plaything?
[1] whatever you want, dear frog,
[-1] she said,
```

Abbildung 1: Wie der Text analysiert wird, Charaktere erkannt werden, sowie Dialogstrukturen und Emotionen. Die Basis für die Generierung der Sprachausgabe

```

i'll dive down and bring your golden ball back to you.
[1] oh, yes,
[-1] she said,
[1] i promise all of that to you if you will just bring the ball back to me.
[-1] but she thought,
? Soundeffect added: +Chorus(delay1:250;amp1:0.54;delay2:400;amp2:-0.10;delay3:200;amp3:0.30)
[1] what is this stupid frog trying to say?
? Soundeffect added: +Chorus(delay1:250;amp1:0.54;delay2:400;amp2:-0.10;delay3:200;amp3:0.30)
[1] he just sits here in the water with his own kind and croaks.
? Soundeffect added: +Chorus(delay1:250;amp1:0.54;delay2:400;amp2:-0.10;delay3:200;amp3:0.30)
[1] he can not be a companion to a human.
[-1] as soon as the frog heard her say "yes" he stuck his head under and dove to the bottom.
[-1] he paddled back up a short time later with the golden ball in his mouth and threw it onto the grass.
[-1] <happy>the princess was filled with joy when she saw her beautiful plaything once again, picked it up, and ran off.
[2] wait, wait,
[-1] called the frog,
[2] take me along.
[2] i can not run as fast as you.
[-1] but what did it help him, that he croaked out after her as loudly as he could?
[-1] she paid no attention to him, but instead hurried home and soon forgot the poor frog, who had to return again to his well.
[-1] the next day the princess was sitting at the table with the king and all the people of the court, and was eating from her golden plate when something came creeping up the marble steps: plip, plop, plip, plop.
[-1] as soon as it reached the top, there came a knock at the door, and a voice called out,
[2] princess, youngest, open the door for me!
[-1] she ran to see who was outside.
[-1] she opened the door, and the frog was sitting there.
[-1] <afraid>frightened, she slammed the door shut and returned to the table.
[-1] the king saw that her heart was pounding and asked,
-----
sender added
ID: 0
Attributes: [BiolFather, Human, Father, Man, Character, Sender, Receiver, Physical]
Voice: EN_ADULT_MALE_A
-----
[0] my child, why are you afraid? is there a giant outside the door who wants to get you?
[1] <afraid>oh, no,
[-1] she answered .
[1] <angry>it is a disgusting frog.
[0] what does the frog want from you?
[1] <sad>oh, father dear, yesterday when i was sitting near the well in the forest and playing, my golden ball fell into the water.
[1] <sad>and because i was crying so much, the frog brought it back, and because he insisted, i promised him that he could be my companion, but i didn't think that he could leave his water.

```

Abbildung 2 Wie der Text analysiert wird, Charaktere erkannt werden, sowie Dialogstrukturen und Emotionen. Die Basis für die Generierung der Sprachausgabe (Fortsetzung von Abbildung 1)

Musterforschung in den Geisteswissenschaften: Werkzeugumgebung zur Musterextraktion aus Filmkostümen

Johanna Barzen¹, Michael Falkenthal¹, Frank Hentschel², Frank Leymann¹

Institut für Architektur von
Anwendungssystemen
Universität Stuttgart
Nachname@iaas.uni-stuttgart.de¹

Musikwissenschaftliches Institut
Universität zu Köln
Frank.Hentschel@uni-koeln.de²

1. Einleitung: Kostümsprache als Mustersprache

In der Literatur zum Filmkostüm findet sich immer wieder der Begriff der „Kostümsprache“ als metaphorische Umschreibung der filmisch vestimentären Kommunikation. Wie diese aber funktioniert, welche Mittel das Kostüm nutzt, um Informationen über die Charaktere, deren Gruppenzugehörigkeit, Stimmungen oder Transformationen, sowie die Zeit- und Ortsgegebenheiten eines Films zu geben, ist nur rudimentär untersucht. Um sich den Funktionsweisen und etablierten Konventionen einer Kostümsprache im Film zu nähern, hat sich das Musterkonzept als fruchtbar erwiesen [SBL12].

Das Konzept des Musters, ursprünglich aus der Architektur stammend [AIS85], hat sich im Besonderen in der Informatik etabliert und findet hier vielseitige Anwendung (Cloud-Computing Patterns, Enterprise Integration Patterns etc.). Definiert wird ein Muster als ein einem vorgegebenen Format folgendes Problem-Lösungspaar, welches eine erprobte Lösung zu einem wiederkehrenden Problem abstrakt erfasst und dieses Wissen so effizient für andere nutzbar macht. Diese Muster werden mit anderen Mustern gleichen Formates untereinander in Beziehung gesetzt, so dass eine Mustersprache entsteht. Im Falle der Filmkostüme ist ein Kostümmuster eine abstrakte Beschreibung einer bewährten Lösung eines wiederkehrenden Designproblems einen adäquaten und schnell verständlichen textilen Ausdruck für beispielsweise eine bestimmte Rolle oder einen Charakterzug zu finden.

Um diese Kostümmuster als abstrakte Lösungsprinzipien (als Essenz vestimentärer Kommunikation) zu entwickeln, müssen erstens die ganzen konkreten Lösungen, in diesem Fall die konkreten Kostüme in Filmen, detailliert erfasst werden [FBB14]. Hierzu haben wir MUSE (MUster Suche und Erkennen) entwickelt. MUSE ist ein Kostümrepository, welches in Sektion 2 näher erläutert wird. Zweitens müssen die erfassten Daten aufbereitet und

ausgewertet werden, um daraus Muster abstrahieren zu können. Wie eine solche Analyse mittels OLAP Cubes aussehen kann wird in Sektion 3 vorgestellt.

2. MUSE: Kostümrepository zur Kostümerfassung

MUSE ist ein, auf die Erfassung von Kostümen spezialisiertes Kostümrepository, das es ermöglicht, Film- und Rolleninformationen, vor allem aber detailgetreue Kostümbeschreibungen einzupflegen. Um eine strukturierte Erfassung und weiterführende Analyse dieser Daten zu ermöglichen, basiert MUSE auf einer umfassenden Bekleidungsontologie, in welche die konkreten Kostüme während des Einpflegens direkt als Instanzen dieser abgelegt werden [Ba13].

Die folgenden Screenshots sollen einen Eindruck vermitteln, wie MUSE die Erfassung von Kostümen unterstützt. Zur Zeit wird hier ein Filmkorpus von 60 Filmen unterschiedlicher Genres eingepflegt, wobei ein Film ca. 200 Kostüme aufweist, welche sich wiederum aus mehreren Basiselementen (Hose, Bluse, etc.) und deren Teilelementen (Ärmel, Kragen, etc.) zusammensetzen.

The screenshot shows a web form titled 'Rolle: Cher Horowitz' with a close button (x) in the top right corner. The form contains several input fields and sections:

- Rolle:** Two text input fields containing 'Cher' and 'Horowitz'.
- Darsteller:** Two text input fields containing 'Alicia' and 'Silverstone'.
- Rollenberuf:** A text input field containing 'Schülerin'.
- Geschlecht:** Three radio buttons labeled 'männlich', 'weiblich' (selected), and 'undefiniert'.
- Dominanter Alterseindruck:** A dropdown menu with 'Jugendlicher' selected and a text input field with '16'.
- Alterseindrücke:** A text input field containing 'Jugendlicher'.
- Dominante Charaktereigenschaften:** A button with a list icon and a text input field containing 'Dominante Charaktereigenschaft'.
- Charaktereigenschaften:** A text area containing a list of adjectives: arrogant, diszipliniert, ehrgeizig, kontaktfreudig, oberflächlich, zickig, angeberisch, aufgedreht, dominant, hochnäsiger, lustig, überdreht, naiv, verspielt, sauerböfisch, aalglatt, abgebrüht, eingebildet, überheblich, unerschrocken, nachdenklich, fröhlich, verführerisch, angsterfüllt, unzufrieden, traurig.
- Familienstand:** A text input field containing 'ledig' with a small 'x' icon to its left.
- Rollenrelevanz:** A dropdown menu with 'Hauptrolle' selected.
- Stereotyp:** A text input field containing 'Zicke, Tussi, Das beliebte Mädchen'.

At the bottom right of the form, there are two buttons: 'Ändern' (blue) and 'Abbrechen' (white).

Screenshot 1: Eingabemaske zur detaillierten Erfassung von Rolleninformationen

Kostümdaten ein-/ausblenden

Basiselemente 5

Neues Basiselement anlegen +

Nur ein Basiselement öffnen

(23) Blazer ▼

Blazer ⓘ ✕

Teilelemente 5

Neues Teilelement anlegen +

Teilelement

(1859) Einreihige Knopfleiste ⓘ	✕
(1743) Hinterteil ⓘ	✕
(1742) Langer Ärmel ⓘ	✕
(39) Revers ⓘ	✕
(1741) Vorderteil ⓘ	✕

(26) Bluse ▶

(27) Anzugweste ▶

(28) Ohrhänger ▶

(1123) Rucksack ▶

Basiselementkomposition 3

Subjekt Objekt +

Subjekt	Operator	Objekt	
(23) Blazer	darüber getragen	(26) Bluse	✕
(23) Blazer	darüber getragen	(27) Anzugweste	✕
(27) Anzugweste	darüber getragen	(26) Bluse	✕

Screenshot 2: Übersicht der Kostümkomposition aus Basis- und Teilelementen und deren Beziehungen zueinander (Operatoren)

Basiselement: Blazer

ID
Init!
Zurücksetzen
×

BasiselementID

Basiselementname ✓

Designs

Formen

Trageweisen

Zustände

Funktionen

Materialien 3

Material	ruck	
	<input type="text"/>	<input style="background-color: #007bff; color: white; border: none; padding: 2px 5px;" type="button" value="+"/>

Materialname	Materialeindruck	
Baumwollstoff	schwer	<input style="background-color: #dc3545; color: white; border: none; padding: 2px 5px;" type="button" value="x"/>
Baumwollstoff	steif	<input style="background-color: #dc3545; color: white; border: none; padding: 2px 5px;" type="button" value="x"/>
Plastik	fest	<input style="background-color: #dc3545; color: white; border: none; padding: 2px 5px;" type="button" value="x"/>

Farben 1

Farbe	Farbeindruck	
<input type="text"/>	<input type="text"/>	<input style="background-color: #007bff; color: white; border: none; padding: 2px 5px;" type="button" value="+"/>

Farbname	Farbeindruck	
Schwarz	kräftig	<input style="background-color: #dc3545; color: white; border: none; padding: 2px 5px;" type="button" value="x"/>

Screenshot 3: Eingabemaske zur Basiselementerfassung (mit aufgeklappter Taxonomie-Eingabehilfe bei Design)

3. Analyse: Auswertung der Kostümdaten

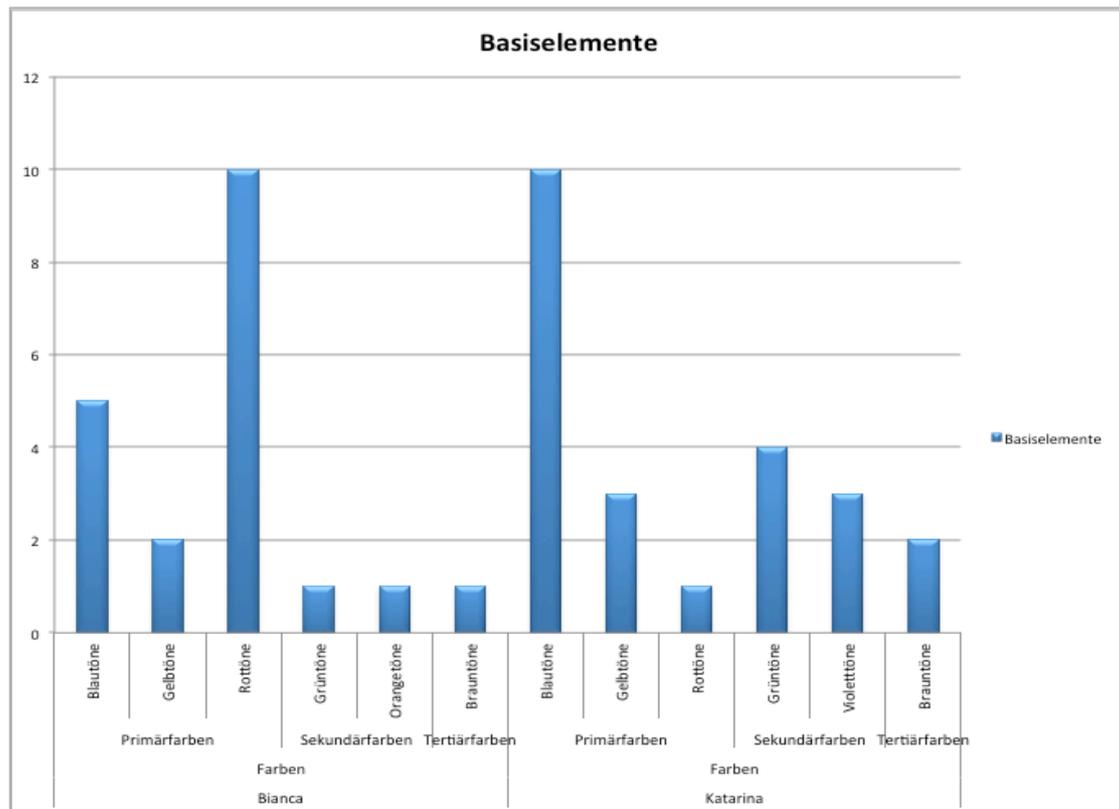
Um die mit MUSE erfassten Daten nutzerfreundlich in ihrem vollen Potential analysieren zu können, haben wir die Werkzeugumgebung so gestaltet, dass die ablaufenden informationstechnischen Auswertungsmethoden so viel Komplexität wie möglich für die Endanwender verbergen. Hierzu werden die Daten mittels eines OLAP Cubes so aufbereitet, dass mit Excel darauf zugegriffen werden kann und hier Auswertungsszenarien definiert werden können, welche die Analyse der Daten aus unterschiedlichen Blickwinkeln in all ihren Dimensionen und Verknüpfungen ermöglicht. Mittels Excel Pivot-Tabellen und als Pivot-Charts visualisiert, kann man sich nun Beispielfragestellungen wie „Welche Farbe ist am häufigsten mit welcher Charaktereigenschaft kombiniert?“, „Ist dieses kostümbildnerabhängig?“, „Werden hochgekremelte Ärmel eher bei passiven oder aktiven Charakteren eingesetzt?“ nähern. Durch das Auftreten von Spitzen in der Häufigkeitsverteilung können dann erste Hinweise auf mögliche Muster gefunden werden.

Screenshot 4 und 5 verdeutlichen, wie man sich beispielsweise dem Einsatz von Farben im Verhältnis zu Charaktereigenschaften nähern kann. Gezeigt wird die Häufigkeitsverteilung der Farben der Kleider der beiden weiblichen Hauptrollen Katarina und Bianca aus „10 Dinge die ich an dir hasse“ (Regie: Junger, 1999).

	A	B	C
1			
2	Originaltitel	10 Things I Hate About You	-Y
3			
4	BE Cube ID Distinct Count	Spaltenbeschriftungen	-Y
5	Zeilenbeschriftungen	Basiselemente	Gesamtergebnis
6	Bianca		15 15
7	▼ Farben		15 15
8	▼ Primärfarben		14 14
9	▶ Blautöne		5 5
10	▶ Gelbtöne		2 2
11	▶ Rottöne		10 10
12	▼ Sekundärfarben		2 2
13	▶ Grüntöne		1 1
14	▶ Orangetöne		1 1
15	▼ Tertiärfarben		1 1
16	▶ Brauntöne		1 1
17	Katarina		18 18
18	▼ Farben		18 18
19	▼ Primärfarben		11 11
20	▶ Blautöne		10 10
21	▶ Gelbtöne		3 3
22	▶ Rottöne		1 1
23	▼ Sekundärfarben		7 7
24	▶ Grüntöne		4 4
25	▶ Violetttöne		3 3
26	▼ Tertiärfarben		2 2
27	▶ Brauntöne		2 2
28	Gesamtergebnis		33 33
29			

Screenshot 4: Pivot-Tabelle

Der unterschiedliche Einsatz der Rot- bzw. Blautöne bei den beiden charakterlich sehr divergierenden Schwestern, lässt bereits erste Rückschlüsse auf Konventionen in deren Einsatz zu. Dies ist allerdings nur als erster Hinweis zu verstehen, der mit weiteren Filmen, Rollen, Charaktereigenschaften, etc. zu überprüfen ist. Genau dabei unterstützt der OLAP Cube.



Screenshot 5: Pivot-Chart zu der Tabelle aus Screenshot 5

4. Ausblick

Zwar ist MUSE, als spezialisiertes Tool zur Kostümerfassung domänenabhängig, die dahinterstehende Methode und das Konzept des Musters zur Wissenserfassung und -repräsentation sind aber auch für die Anwendung in anderen Bereichen der Geisteswissenschaften ein vielversprechender Ansatz und gehen weit über die Kostümforschung hinaus [BL14].

Angedacht ist der Einsatz zur Extraktion von musikalischen Mustern, um Charakteristika und Topoi musikalischer Artefakte, die sich mit den herkömmlichen musikwissenschaftlichen Konzepten wie „Thema“, „Motiv“ oder „Stil“ nicht erfassen lassen, herauszuarbeiten und eventuell im Hinblick auf ihre semantische oder expressive Funktion deuten zu können.

5. Referenzen

- [AIS85] Alexander, C.; Ishikawa, S.; Silverstein, M.; Jacobson, M.; Fiksdahl-King, I.; Angel, S.: A Pattern Language: Towns, Buildings, Constructions. Oxford University Press, 1977.
- [Ba13] Barzen, J.: Taxonomien kostümrelevanter Parameter: Annäherung an eine Ontologisierung der Domäne des Filmkostüms, Universität Stuttgart, Technischer Bericht Nr. 2013/04, 2013.
- [BL14] Barzen, Johanna; Leymann, Frank: Kostümsprache als Mustersprache: Vom analytischen Wert Formaler Sprachen und Muster in den Filmwissenschaften, In: DHd 2014.
- [FBB14] Falkenthal, M.; Barzen, J.; Breitenbücher, B.; Fehling, C.; Leymann, F.: From Pattern Languages to Solution Implementations. In: Proceedings of the 6th International Conference on Pervasive Patterns and Applications, Venice, 2014.
- [SBL12] Schumm, D.; Barzen, J.; Leymann, F.; Ellrich, L.: A Pattern Language for Costumes in Films. In: Proceedings of the 17th European Conference on Pattern Languages of Programs (EuroPLOP), Irsee, 2012. ACM Press, New York, 2012.

Anforderungen und Bedürfnisse von Geisteswissenschaftlern an einen digital gestützten Forschungsprozess

Oona Leganovic, Viola Schmitt, Juliane Stiller, Klaus Thoden & Dirk Wintergrün
Max-Planck-Institut für Wissenschaftsgeschichte

Cluster 1 von DARIAH-DE¹ (Wissenschaftliche Begleitforschung) hat zum Ziel den geisteswissenschaftlichen Forschungsprozess zu analysieren um Bedürfnisse von Fachwissenschaftlern im Hinblick auf virtuelle Forschungsinfrastrukturen besser zu verstehen. Durch diese Arbeit sollen die innerhalb von DARIAH-DE entwickelten Dienstleistungen an die fachwissenschaftlichen Anforderungen angepasst werden. Um dies zu verwirklichen hat sich Cluster 1 die folgende drei Schritte vorgenommen:

1. Analyse der Beziehung zwischen geisteswissenschaftlichen Forschungsprozessen und den von digitalen Tools abgedeckten Prozessen,
2. Kartierung bisher genutzter digitaler Tools und Methoden um eventuelle Lücken aufzudecken,
3. Formulierung von Anforderungen an virtuelle Forschungsumgebungen zur Unterstützung des geisteswissenschaftlichen Forschungsprozesses.

Dieses Poster wird die Ergebnisse des ersten Arbeitsschrittes darstellen.

Es wurden vorhandene Modelle, die den geisteswissenschaftlichen Forschungsprozess konzeptionell erfassen auf ihre Gemeinsamkeiten und Unterschiede hin untersucht. Darauf aufbauend wurde ein auf unsere Bedürfnisse zugeschnittener Forschungskreislauf modelliert, der sowohl digitale als auch klassische Forschungsprozesse einbeziehen soll.

Es gibt eine Vielzahl von Modellen, die den Forschungsprozess vereinfacht darstellen und ihn auf Konzepte oder Aktivitäten reduzieren. Die Grundlage all dieser Überlegungen hat Unsworth mit seinen Primitiven gelegt (Unsworth, 2000). In eine ähnliche Richtung geht TADiRAH (Taxonomy of Digital Research Activities in the Humanities) - eine Taxonomie geisteswissenschaftlicher Forschungsmethoden und -ziele (Borek et al., 2014). Auch Bernardou und andere (2010) haben ein Modell entwickelt, das den geisteswissenschaftlichen Forschungsprozess abbildet mit besonderem Augenmerk auf die Ziele der beschriebenen Aktivitäten. Innerhalb des EU-geförderten Projektes DM2E² (Digital Manuscripts to Europeana) wurde in einem der Meilensteine das Scholarly Domain Model (SDM) beschrieben (Gradmann & Henicke, 2012).

Wir haben die Modelle Unsworth's Primitives, TaDiRAH und das SDM aufeinander abgebildet um Gemeinsamkeiten und Unterschiede festzustellen. Abbildung 1 zeigt die Primitiven von Unsworth (2000), deren Verhältnis zu den Primitiven und Aktivitäten des SDM und der TaDiRAH Taxonomie. Man kann gut erkennen, dass es viele Überschneidungen, auch in der Terminologie, gibt. Weiterhin herrscht Einigkeit über die Aktivitäten, die während des geisteswissenschaftlichen Forschungsprozesses stattfinden. Auffällig ist, dass die Aktivitäten sich natürlich unterscheiden im Hinblick auf den Teil des Prozesses, den sie abbilden. So ist

¹ <https://de.dariah.eu/>

² <http://dm2e.eu/>

TaDiRAH sehr auf die Abbildung digitaler Arbeitsprozesse, die mit Software erledigt werden fokussiert und hat deswegen eine Aktivität "Storage". Dies spielt in den anderen Taxonomien eine untergeordnete Rolle und ist oft Teil von anderen Aktivitäten, wie "Aggregation" beim SDM und "Sampling" bei Unsworth.

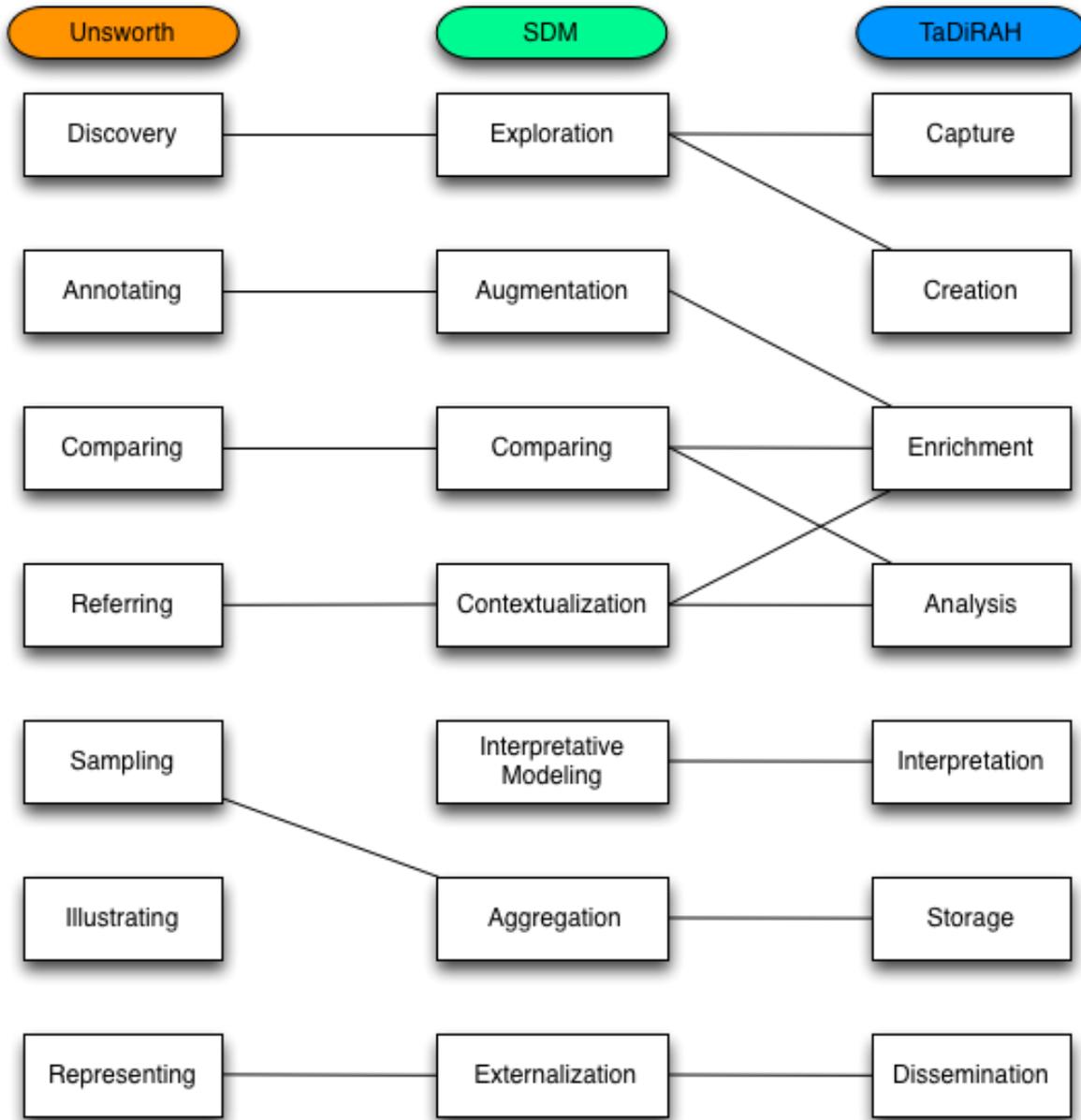


Abbildung 1: Abbilden der Unsworth'schen Primitiven, des Scholarly Domain Models und TaDiRAH

Die vorangegangenen Modelle versuchen auf Basis der vorhandenen Tools Kategorisierungen zu erzielen oder auf Basis des Forschungsprozesses Arbeitsabläufe zu konzeptualisieren. Wir wollen anhand des geisteswissenschaftlichen Forschungsprozesses darstellen, was Tools

leisten müssen um diesen zu unterstützen. Ziel ist es Lücken aufzudecken und zu verstehen, wo digitale Dienstleistungen den Forschungsprozess besser unterstützen können und müssen.

Als ein Schritt zu diesem Ziel untersuchen wir, wie dieser prototypische Forschungsprozess sich in einem digitalen Arbeitsablauf abbilden lässt und wo sich Lücken befinden und wodurch diese entstehen. Dafür haben wir uns auf Grundlage der oben beschriebenen Modelle vor allem auf die Ergebnisse (und den Output) der verschiedenen Aktivitäten konzentriert (Abbildung 2).

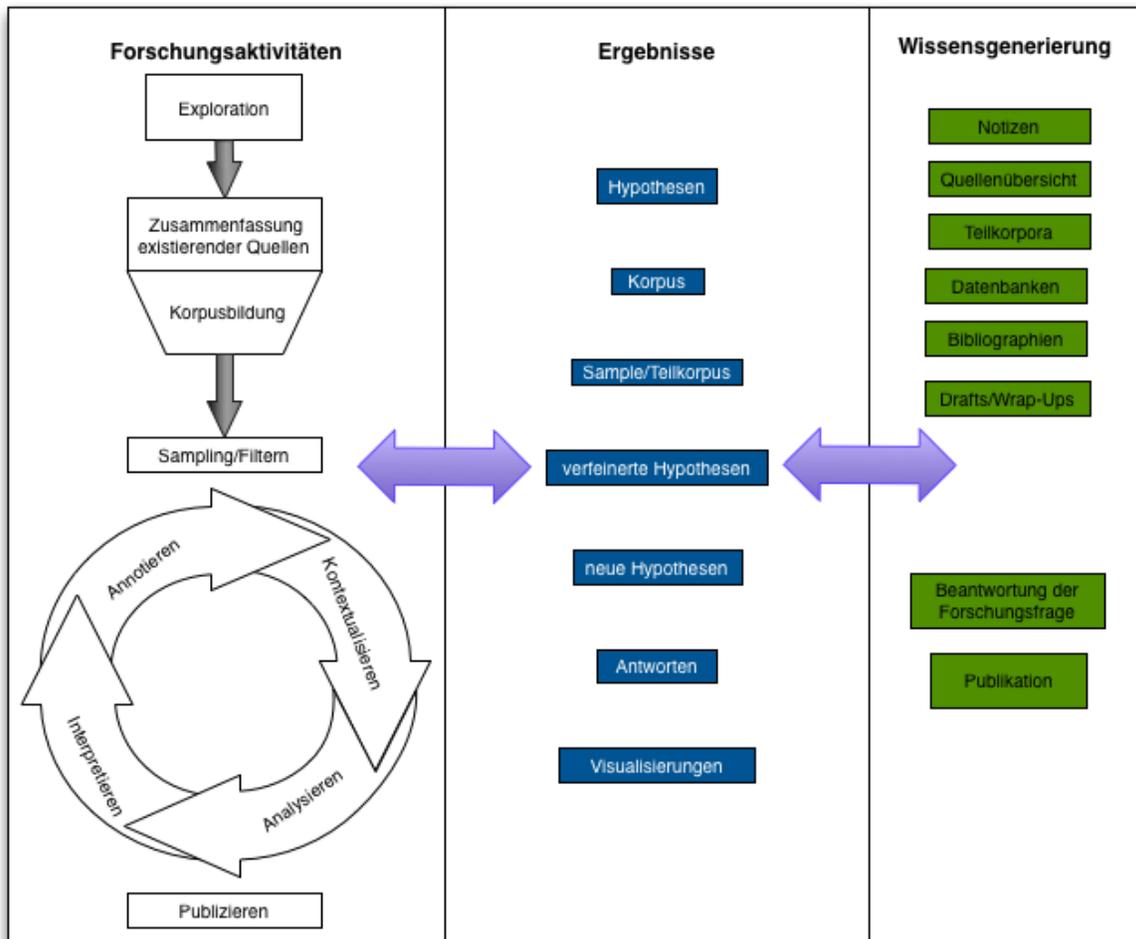


Abbildung 2: Forschungsaktivitäten, deren Ergebnisse und Output als Wissensgenerierung.

Es wurden die wichtigsten Ergebnisse, die in einem digitalen Prozess gespeichert und weiterverarbeitet werden, gelistet. Dabei ist zwischen den Zwischenergebnissen jedes einzelnen Schritts (Spalte 2 Abbildung 2) und dem Output, der in seiner gegenständlichen Form in der nächsten Aktivität verarbeitet wird (Spalte 3 Abbildung 2), zu unterscheiden. Dieses generierte Wissen kann mit anderen einzelnen Forschern aber auch der Öffentlichkeit geteilt werden. Dies kann die Publikation sein, die klassischerweise am Ende des Forschungsprozesses steht, aber auch Quellenübersichten, Datenbanken oder Bibliographien, die vor der Veröffentlichung angelegt werden. Häufig kann der Forschungsprozess nicht eindeutig modelliert und mit Aussagen versehen werden, die starr vorgeben, dass bestimmte Aktivitäten immer zu einer bestimmten Form des Ergebnisses und der Wissensgenerierung führen. Trotzdem ist es durchaus sinnvoll sich im Bezug auf die digitale Unterstützung des

Forschungsprozesses deutlich zu machen, dass jede Aktivität eine Form des Outputs produziert, auf dem der Forscher idealerweise im nächsten Schritt seines Denkprozesses aufbauen möchte. Dies gilt vor allem für den digitalen Arbeitsverlauf; lästiges hin- und her Kopieren und Konvertieren in verschiedene Datenformate beim Wechsel von Tools ist eines der Probleme, die Brüche im digitalen Forschungsprozess hervorrufen.

Mit diesem Poster möchten wir unsere Überlegungen vorstellen und zur Diskussion einladen, wie die Bedürfnisse von Fachwissenschaftlern in virtuellen Forschungsumgebungen Berücksichtigung finden und digitale Dienstleistungen aufgebaut werden können, die Geisteswissenschaftler in ihrer Arbeit unterstützen.

Literatur

Benardou, Agiatis; Constantopoulos, Panos; Dallas, Costis; Gavrilis, Dimitris (2010): A Conceptual Model for Scholarly Research Activity. iConference 2010. Online: <https://www.ideals.illinois.edu/handle/2142/14945>

Borek, Luise; Quinn Dombrowski; Matthew Munson; Jody Perkins; Christof Schöch (2014): Scholarly primitives revisited: towards a practical taxonomy of digital humanities research activities and objects, Digital Humanities Conference 2014, Lausanne, Switzerland, July 7-12, 2014

Gradmann, Stefan; Hennicke, Steffen (2012): Intermediary Research Report on DH Scholarly Primitives (MS 3). Project DM2E.

Unsworth, J. (2000, May). Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this? Symposium on Humanities computing: Formal methods, experimental practice, King's College, London, Online: <http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html>

,Was heißt und zu welchem Ende produziert man ein geisteswissenschaftliches E-Journal?'

Innovationspotentiale des digitalen Publizierens am Beispiel der *Zeitschrift für Digital Humanities* (ZfDH)

Constanze Baum (Wolfenbüttel), Timo Steyer (Wolfenbüttel)

Die *Digital Humanities* sind dabei, die geistes- und kulturwissenschaftliche Forschung grundlegend zu verändern. Durch die Inhalte und Methoden der *Digital Humanities* werden aber nicht nur neue Zugangswege, Fragen und Auswertungsmöglichkeiten zu bzw. an Primärquellen ermöglicht, sondern es eröffnen sich auch für die Präsentation und Publikation von Forschungsdaten und -ergebnissen innovative Alternativen zu traditionellen Printmedien. Die *Zeitschrift für Digital Humanities* (ZfDH) wird diese beiden Felder miteinander kombinieren, indem sie als dezidiertes Organ für die *Digital Humanities* im deutschsprachigen Raum nicht nur Themen der *Digital Humanities* veröffentlicht, sondern selbst ein Produkt der *Digital Humanities* ist: Hier werden neue Verfahren und Methoden digitalen Publizierens im Sinne einer Prototypentwicklung eines E-Journals für die Geisteswissenschaften ausgelotet. Das Poster wird sowohl die Innovationspotentiale des E-Journals zur Diskussion stellen, als auch über den gegenwärtigen Stand des Projektes informieren. Insofern scheint es gerechtfertigt, in Anlehnung an Schillers Antrittsrede vor 125 Jahren – „Was heißt und zu welchem Ende studiert man Universalgeschichte?“ (1789) – programmatisch und grundsätzlich über Wege und Potentiale eines E-Journals im Bereich der Geisteswissenschaften nachzudenken und zu fragen: Was heißt und zu welchem Ende produziert man ein geisteswissenschaftliches E-Journal?

Es werden dabei Felder des digitalen Publizierens aufgezeigt, die die Bereiche der Beitragsakquise ebenso wie einen digital grundierten Workflow, ein offeneres Review-Verfahren und die vielversprechenden Möglichkeiten der E-Distribution der Zeitschrift betreffen. In den Geisteswissenschaften fehlen hier auf vielen Feldern noch Standards und Normen für E-Journale. Insofern versteht sich die ZfDH als Pilotprojekt und Innovationsgeber. In Anlehnung und Abgrenzung zu Projekten aus den Natur- und Technikwissenschaften sollen daher die Potentiale herausgearbeitet werden, die das digitale Publizieren in den Geisteswissenschaften haben kann. Denn im Bereich der Softwareentwicklung ist trotz verschiedener vorhandener Programme (*Open Journal System*) der Innovationsgrad längst nicht auf dem Niveau, wie er auf anderen Feldern der *Digital Humanities* bereits erreicht worden ist. Die Entwicklung der *Zeitschrift für Digital Humanities* beinhaltet daher sowohl die Ausarbeitung eines innovativen Workflows als auch dessen konkrete technische Umsetzung. Ausgegangen wird hierbei nicht von einer fertigen Softwarelösung, vielmehr wird die Software entsprechend den formulierten Anforderungen an das E-Journal modular entwickelt und aufgebaut.

Innovation im Bereich von wissenschaftlich orientierten E-Journals besteht vor allem in der freien Verfügbarkeit der Inhalte (OA) und der wesentlich schnelleren Publikation der Artikel, ohne dabei auf eine umfangreiche Qualitätskontrolle zu verzichten. Gedacht ist zurzeit an ein transparentes Review-

Verfahren, das es dem Autor ermöglichen wird, das jeweilige Gutachten einzusehen und darauf zu reagieren und bei Bedarf eine revidierte Fassung einzureichen sowie eine Gesamtbeurteilung der Gutachter für alle Nutzer öffentlich zu machen. Alle Fassungen bleiben mittels eindeutiger DOI-Nummern recherchier- und archivierbar. Vorab steht die Entscheidung, die redaktionsgeprüfte Erstfassung eines Artikels bereits nach einer ersten Routine online zu stellen. Die Qualitätskontrolle ist demzufolge im Sinne einer Liberalisierung von Wissensdiskursen (*Open Science*) als moderiertes *post-publication-peer-review*-Verfahren angedacht.

Um die Nachnutzung der Artikel zu gewährleisten, werden die Artikel unter einer freien Lizenz veröffentlicht und in XML bereitgestellt. Ob TEI sich auch als Standard für wissenschaftliche Sekundärliteratur im E-Journalbereich eignet, ist eine der zentrale Forschungsfragen des Projektes. Innovationspotentiale bestehen auch im Bereich weiterer Serviceleistungen, die Printmedien nicht bieten können, dazu zählen semantische Anreicherungen wie ein weitreichendes Verlinkungssystem, die Einbindung bestehender Normdaten und die Distribution der Zeitschrift über standardisierte Schnittstellen (Katalogisierung, Indexierung). Ein implementiertes Metriksystem liefert Angaben über die wissenschaftliche Nachnutzung einzelner Artikel. Die Artikel des E-Journals werden periodisch in Ausgaben zusammengefasst, dies dient vor allem der Erschließung und Distribution. Umfangreiche Suchfunktionen, Verschlagwortungen, Metadaten und Rubrizierungen bieten weitere digitale Möglichkeiten der Erschließung der Artikel, die parallel dazu zur Verfügung gestellt werden.

Darüberhinaus eröffnen sich für digitale Publikationen über den Text hinaus weitreichende Optionen für die Einbettung digitaler Medien, seien es Bilder, Videos, Tondokumente, Blogbeiträge oder Twitterfeeds. Die (dynamische) Aggregation unterschiedlicher Ressourcen bringt die Frage auf, wie eine persistente Identifizierung der einzelnen Bestandteile möglich sein wird und welche wissenschaftliche Relevanz diesem Quellenmaterial in der Forschung künftig zugewiesen wird. Es stellt sich demnach auch die Frage, inwieweit durch solche Formen digitalen Publizierens neue Inhalte für die wissenschaftliche Beschäftigung erschlossen werden können.

Wittgensteins Nachlass: Aufbau und Demonstration der FinderApp WiTTFind und ihrer Komponenten

Yuliya Kalasouskaya, Matthias Lindinger, Stefan Schweter, Roman Capsamun
Y.Kalasouskaya1@campus.lmu.de, matthias.lindinger@campus.lmu.de,
Stefan.Schweter@campus.lmu.de, r.capsamun@campus.lmu.de
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München

1 EINLEITUNG

Das von uns erstellte Poster soll den Aufbau und Einsatz der FinderApp WiTTFind mit den zugehörigen WAST-Tools¹ als *open source* Tool vorstellen. Im Mittelpunkt stehen die optimierte Browseroberfläche, zugrunde liegende Texte der FinderApp, Faksimile mit OCR, Faksimile Reader und den Einsatz des Finders als open source Programm. Für Interessierte werden wir die FinderApp vorführen.

2 NEUERUNGEN DER FINDERAPP

Seit 2 Jahren wird mit unserem Finder in der Nachlassforschung von Ludwig Wittgenstein gearbeitet und die von uns entwickelten Programme werden stetig optimiert. Eine zusätzliche Motivation für die Weiterentwicklung und Erweiterung war auch die Verleihung des EU-AWARDS 2014, der vom EU Projekt Digitised Manuscripts to Europeana (DM2E) ausgeschrieben wurde. Neuerungen des Finders bestehen darin, dass die Weboberfläche optimiert wurde, mehrere Dokumente parallel durchsucht werden können und eine lemmatisierte symmetrische Vorschlagsuche sowie ein neuer Faksimile-Reader integriert wurden. Die wichtigste Neuerung bei unserem Finder ist jedoch, dass WiTTFind für andere Forschungsprojekte geöffnet wurde und als open source für andere Projekte der Digital Humanities einsetzbar sein wird. Die Applikation ist unter dem folgenden Permalink zu erreichen:

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

3 DIE KOMPONENTEN DES FINDERS

3.1 BENUTZEROBERFLÄCHE

Als erstes wollen auf dem Poster wir die Gestaltung der neuen benutzerfreundlichen und interaktiven Hauptseite der FinderApp vorstellen. Um die Bedienung der Anwendung übersichtlich zu gestalten, werden den Nutzern verschiedene Suchumgebungen angeboten. In der neuen Version können mehrere Text-Ressourcen parallel durchsucht werden, weshalb die Applikation um eine *multidoc*-Struktur erweitert wurde. Die Darstellung der Treffer wird auf die ausgewählten Dokumente beschränkt. In der nächsten Abbildung ein Beispiel zur *multidoc* Oberfläche:



Die Antwort auf die Benutzeranfrage wird mit Hilfe des *LocalStorage* Konzepts im Browser gespeichert. Bearbeitet werden die Daten und die *multidoc* Funktionen nur auf der Client Seite (Web-Browser) ohne Server-Zugriff. Zur Interaktivität und Lebendigkeit der Seite tragen die modernen Techniken von *JQuery* und *HTML5* bei.

¹ Wittgenstein Advanced Search Tools

3.2 FAKSIMILE-READER

Das zweite Thema des Posters stellt den Faksimile-Reader (siehe Bild:Faksimile-Reader) vor, der es erlaubt komplementär durch die Faksimile der Edition zu blättern und gleichzeitig die gefundenen Textstellen im Bild hervorzuheben. Dieser ist in Javascript sowie den Bibliotheken *jQuery* und *turn.js* programmiert. Zum *Highlighting* der einzelnen Treffer wird eine Liste von Koordinaten verwendet, die im Javascript-eigenen JSON-Format vorliegt. Zur schnellen Darstellung der Faksimile werden immer nur die Seiten geladen, die der Anfrage des Benutzers entsprechen. Somit kann der Anwendungsnutzer das komplette Dokument in Faksimile-Form durchblättern. Weitere Features des Readers sind eine dynamische Anpassung der Ansicht an das Browserfenster und eine kurze Bedienungsanleitung, die beim Start angezeigt wird. Damit die Faksimile zusammen mit den gefundenen, farblich hervorgehobenen Treffern dargestellt werden können, müssen die Faksimile mit der open source OCR Software *tesseract* verarbeitet werden. Je nach Qualität der Faksimile müssen die extrahierten OCR-Ergebnisse manuell nachbearbeitet werden. Dazu haben wir spezielle Tools entwickelt.

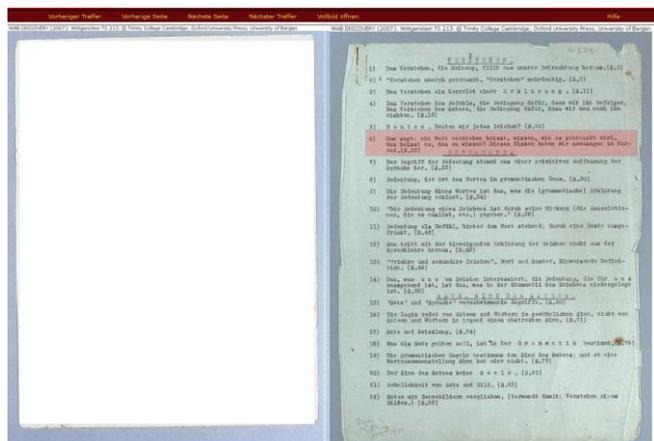


Bild:Faksimile-Reader

² Laboratoire d'Automatique Documentaire et Linguistique, Paris

4 UNSERE FINDERAPP FÜR ANDERE DIGITAL HUMANITIES PROJEKTE

In einem weiteren Thema des Posters geht es um eins der wichtigsten Ziele unseres Projekts: Die FinderApp und die WAST-Tools sollen plattformunabhängig einem breiten Forschungskreis zur Verfügung stehen.

4.1 DIE TEXTE DER EDITION

Die FinderApp findet Wörter, semantische Begriffe und Satzphrasen über mehrere Dokumente hinweg, sofern die Dokumente in dem XML-TEI-P5 Format vorliegen. Dieses XML-Format wird von uns CISWAB genannt und in einer eigenen *Document Type Definition (DTD)* beschrieben.

4.2 ELEKTRONISCHES VOLLFORMEN-LEXIKON

Zu den Texten einer Edition benötigt die FinderApp ein elektronisches Lexikon im DELA Format². Das CIS verfügt über das größte deutsche Vollformenlexikon, das bei der Entwicklung eines eigenen „Editionslexikons“ herangezogen werden kann.

4.3 SOFTWARE-KOMPATIBILITÄT UNSERER FINDERAPP

Da bei unseren Programmen eine große Anzahl unterschiedlicher Programmiersprachen und Libraries im Einsatz sind, die von Standarddistributionen abweichen, setzen wir die quelloffene Containervirtualisierungssoftware namens *docker* ein. Bei dieser Technologie werden alle von WiTTFind benötigten Programme, Module und Libraries in einem Softwarecontainer zusammengefasst. Jeder Anwender, der auf seinem Rechner die *docker* Software³ installiert hat, kann unseren Finder mit WAST-Tools lokal auf seinem Rechner einsetzen. Zur Entwicklungsverwaltung verwenden wir das

³ <https://www.docker.com/>

Versionsverwaltungsprogramms Git und das web-basierte Versionsverwaltungs-Management-Werkzeug GitLab.

Am Posterstand werden wir auf Laptops mit verschiedenen Betriebssystemen unsere FinderApp WiTTFind und WAST-Tools vorstellen.

5 PUBLIKATION UND AWARD

EU AWARD: <http://dm2e.eu/open-humanities-awards-round-2-winners-announced/>

Max Hadersbeck, Alois Pichler, Florian Fink, Øyvind Liland Gjesdal: Wittgenstein's Nachlass: WiTTFind and Wittgenstein advanced search tools (WAST). Madrid, DATECH 2014: 91-96 <http://wast.cis.uni-muenchen.de/tutorial>

Computerlinguistische Verfahren zur Aufdeckung struktureller Ähnlichkeiten in Narrativen

Einführung

In diesem Beitrag stellen wir eine Methode zur automatischen Erkennung von strukturellen Ähnlichkeiten narrativer Texte auf der Handlungsebene vor. Dafür operationalisieren wir strukturelle Ähnlichkeiten als (intertextuelle) Verbindungen (*Alignments*) zwischen Ereignissen. Die verwendeten Alignierungsalgorithmen bauen auf automatisch erzeugten linguistischen Analysen der Texte auf und verwenden als Kriterien Eigenschaften verschiedener linguistischer Ebenen. Ziel unseres Ansatzes ist es, materiell in Texten vorliegende Ähnlichkeiten auffindbar zu machen und hervorzuheben, so dass sie von Wissenschaftlerinnen und Wissenschaftlern zielgerichtet analysiert und interpretiert werden können.

Anwendungsszenarien

Die Untersuchung struktureller Ähnlichkeiten zwischen Narrativen spielt in vielen geisteswissenschaftlichen Disziplinen eine Rolle. Als Beispielszenarien verwenden wir die Märchen- und Ritualforschung.

Ähnlichkeiten zwischen **Märchen** sind auf verschiedenen Granularitätsebenen untersucht worden. Propp (1958) veröffentlichte eine Analyse, in der in russischen Märchen prototypische Handlungen und Charaktere identifiziert werden.

Regelmäßigkeiten im Auftreten von Handlungen und Charakteren werden in einer sog. „Morphology of the Folktale“ erfasst. Damit sollen typische Handlungsmuster (Ereignis X folgt auf Ereignis Y) beschrieben werden. Am anderen Ende der Granularitätsskala existieren Sammlungen wie der ATU-Index (Uther, 2014), in dem Märchen mit gleichen Handlungselementen (Aussetzen von Kindern) oder Charakteren (Lebkuchenhaus) in Klassen zusammengefasst werden.

Im Bereich der **Ritualforschung** werden Rituale aus diversen religiösen, kulturellen oder politischen Kontexten untersucht. Unter dem Stichwort „Ritualgrammatik“ (vgl. Hellwig und Michaels, 2013) wird diskutiert, dass in verschiedenen Ritualen ähnliche Handlungen vorkommen und Teilnehmer ähnliche Rollen übernehmen.

Verschiedene Forscher vertreten die Auffassung, dass die Zusammensetzung wiederkehrender Ereignisse zu Ritualen Regeln folgt. Existierende Überlegungen zur Ritualgrammatik sind nicht formalisiert und daher für eine automatische Analyse nur begrenzt nutzbar.

Um unsere Methode entwickeln und testen zu können, haben wir für diese beiden Szenarien ein englischsprachiges Korpus zusammengestellt, das mehrere Beschreibungen des gleichen Typs enthält (ATU-Märchenklasse bzw. Ritualtyp).

Computerlinguistische Verarbeitung

Wir wenden die gleichen computerlinguistischen Komponenten auf beide Korpora an. Damit werden linguistische Repräsentationen für Wortarten, (syntaktische) Abhängigkeitsrelationen, semantische Rollen, Wortbedeutungen und Koreferenzketten erstellt. Verknüpft ergeben diese Annotationen eine Diskursrepräsentation, die als Basis für die Alignierungsverfahren verwendet wird. Da Ritualbeschreibungen untypische linguistische Phänomene enthalten, wurden sämtliche Komponenten auf die Domäne angepasst (*Domain Adaptation*). Dadurch konnten deutliche Qualitätssteigerungen der computerlinguistischen Analyse erreicht werden.

Alignierungsexperimente

Drei Alignierungsalgorithmen mit unterschiedlicher Mächtigkeit wurden verglichen: *Sequence alignment* (Needleman-Wunsch, 1970) ist der einfachste Algorithmus, der ausschließlich paarweise und nicht-kreuzende Alignierungen erzeugen kann. *Graph-based predicate alignment* (GPA; Roth, 2014, Roth & Frank, 2012) kann paarweise und kreuzende Alignierungen erzeugen. *Bayesian model merging* (BMM; Stolcke & Omohundro, 1993) ist der mächtigste Algorithmus, der Alignierungen beliebiger Länge mit Überkreuzungen erzeugen kann. Diese drei Algorithmen wurden in zwei Experimenten evaluiert: In einer intrinsischen Evaluation wurden die Ergebnisse mit einem von zwei Ritualwissenschaftlern parallel erzeugten Goldstandard verglichen ($\kappa=0.61$). Dabei erzielte BMM die besten Ergebnisse insgesamt und GPA die besten Ergebnisse auf einem Einzeldokumentpaar.

Im zweiten Experiment wurde aus den automatisch erzeugten Alignierungen ein Maß für Dokumentenähnlichkeit berechnet und in einem Clustering-Verfahren eingesetzt. Das Ergebnis des Clusterings – eine Einteilung der Dokumente auf Basis der errechneten strukturellen Ähnlichkeit – konnte dann mit der Gruppierung verglichen werden, die „natürlicherweise“ in den Korpora vorkommt (Ritualtypen bzw. ATU-Klassen). Dabei zeigten sich wieder GPA und BMM als die leistungsstärksten Algorithmen.

Visualisierung und Nutzung

Um es Wissenschaftlerinnen und Wissenschaftlern aus der Ritual- bzw. Märchenforschung zu ermöglichen die Analysen zu nutzen, haben wir Visualisierungen entwickelt, die eine systematische Untersuchung der gefundenen Ähnlichkeiten ermöglichen. Auf einer Vogelperspektive stellen wir die Dokumentenähnlichkeit in einer Heatmap dar. Auf interessante, dicht verknüpfte Stellen können wir hinweisen, indem für jedes Ereignis ein *connectivity score* in einem Diagramm angezeigt wird. Eine detaillierte Darstellung der Einzelereignisse (mit Teilnehmern und Kontext-Ereignissen) ist ebenfalls möglich. Direkt aus der Diskursrepräsentation können wir außerdem eine Visualisierung des sozialen Netzwerks erzeugen, in der wichtige Entitäten (Charaktere, Gegenstände und Materialien) in einem Netzwerk angezeigt und gemeinsam auftretende Figuren verknüpft werden.

Konklusion

Der Posterbeitrag präsentiert eine Methode zur Erkennung struktureller Ähnlichkeiten zwischen narrativen Texten. Die Ähnlichkeiten werden basierend auf computerlinguistischen Analysen vollautomatisch identifiziert und können zielgerichtet auf unterschiedlichen Granularitätsebenen dargestellt und manuell inspiziert werden. Damit eignet sich die Methode auch zur Analyse von größeren Datenmengen, ohne bestimmte Interpretationen vorwegzunehmen. Eine ausführliche Darstellung des Verfahrens sowie des geisteswissenschaftlichen Anwendungskontexts findet sich in Reiter (2014) und Reiter et al. (2014). Auf einer methodischen Ebene zeigt sich in diesem Projekt, dass komplexe linguistische Analysen auch für nicht-kanonische Textsorten erstellt werden können und eine vielversprechende Ausgangsbasis für Analysen darstellen. Die Besonderheiten natürlicher Sprache (z.B. Ambiguität, Vielseitigkeit) stellen für automatische Verarbeitung eine große Herausforderung dar, werden aber in der Computerlinguistik bereits untersucht. Auf (computer-)linguistische Analysen aufzubauen erlaubt die Untersuchung komplexer semantischer Phänomene, die vergleichsweise eng mit den Zielkategorien vieler Geisteswissenschaften verwandt sind.

Bibliographie

Oliver Hellwig and Axel Michaels. Ritualgrammatik. In Christiane Brosius, Axel Michaels, and Paula Schrode, Hrsg., *Ritual und Ritualdynamik*, S. 144–150. Vandenhoeck & Ruprecht, Göttingen, Germany, 2013.

Saul B. Needleman and Christian D. Wunsch. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. *Journal of Molecular Biology*, 48(3):443–453, March 1970.

Vladimir Yakovlevich Propp. *Morphology of the Folktale*. University of Texas Press, Austin, TX, 2nd edition, 1958. Translated by Laurence Scott (Original work published 1928).

Nils Reiter. *Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms*. PhD thesis, Heidelberg University, June 2014.

Nils Reiter, Anette Frank, and Oliver Hellwig. An NLP-based cross-document approach to narrative structure discovery. *Literary and Linguistic Computing*, 29(4):583–605, 2014.

Michael Roth. *Inducing Implicit Arguments via Cross-document Alignment – A Framework and its Applications*. PhD thesis, Heidelberg University, 2014.

Michael Roth and Anette Frank. Aligning predicates across monolingual comparable texts using graph-based clustering. In Jun'ichi Tsujii, James Henderson, and Marius

Paşca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 171–182, Jeju Island, Korea, July 2012.

Andreas Stolcke and Stephen Omohundro. Hidden markov model induction by bayesian model merging. In Steve J. Hanson, J. D. Jack D. Cowan, and C. Lee Giles, Hrsg., *Advances in Neural Information Processing Systems*, volume 5, pages 11–18. Morgan Kaufmann, San Mateo, California, 1993.

Hans-Jörg Uther. *The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*. Number 284–286 in FF Communications. Suomalainen Tiedeakatemia, Helsinki, 2004.

Netzwerke sehen

Matej Ďurčo, ACDH-ÖAW

In diesem Beitrag stellen wir eine Webapplikation zur Visualisierung und interaktiven Erkundung von Graphen und Netzwerken vor. Die Applikation ist ursprünglich im Kontext der Forschungsinfrastruktur CLARIN entstanden, mit dem Ziel die komplexe Datendomäne der Metadaten-Profile der Component Metadata Infrastructure¹ (CMDI) (Broeder et al., 2010) besser fassbar zu machen. (Ďurčo, 2013). Im Laufe der Entwicklung hat sich diese Applikation zu einem generischen Viewer für jede Art von graph-basierten Daten weiterentwickelt.

Die Applikation kann auch im Vergleich mit alternativen weit verbreiteten Tools bestehen. Gephi² bietet zwar wesentlich mehr Funktionalität zum automatischen Analysieren von Graphen, ist aber eine Client-Applikation, die lokal installiert wird, und die Möglichkeiten der dynamischen Navigation im Graphen sind auch nicht so reichhaltig, wie in der vorgestellten Applikation. Die traditionelle command-line Applikation GraphViz³ ist zwar sehr stark im eleganten Layoutieren der Graphen, ist aber eine rein statische Anwendung ohne graphisches User Interface, bietet also keine Möglichkeit interaktiv zu arbeiten.

Die Applikation basiert auf der open-source javascript Bibliothek d3⁴ und läuft nach dem anfänglichen Laden vollständig client-seitig. Es bietet mehrere miteinander verknüpfte Ansichten und eine Reihe von Optionen zum Manipulieren der dargestellten Graphen. So ist es möglich mehrere Knoten auszuwählen und sich aus dem zugrundeliegenden geordneten Graphen beliebig viele Ebenen von Vorgänger- bzw. Nachfolgerknoten anzeigen zu lassen. Ebenfalls werden mehrere vordefinierte Layout-Algorithmen angeboten. Das Layout kann die Stärke der Verbindungen reflektieren, ebenso kann die Größe und Farbe der Knoten verwendet werden, um weitere Dimensionen visuell zu kodieren. Der aktuell angezeigte Graph, kann entweder als Link verschickt oder als SVG-Grafik exportiert und weiter verarbeitet werden.

Daten

Neben den CMDI Metadaten, für welche die Applikation ursprünglich vorgesehen war, wurde die Applikation bereits erfolgreich mit ganz anderen Datensätzen erprobt. So wurde zum Beispiel das „Philosophen-Influenz-Netzwerk“ visualisiert. Dafür wurden die Daten aus der dbpedia über den SPARQL-Endpoint abgefragt (Philosophen und ihre *influenced/influencedBy* Beziehungen), diese wurden durch eine einfache XSLT-Transformation in das Eingabeformat umgewandelt und in die Applikation importiert. Mit minimalem Aufwand konnte dergestalt eine große Datenmenge wesentlich besser erfasst und erkundet werden als dies mit konventionellen Instrumenten möglich wäre. (siehe Abb. 1)

Es ist vorgesehen, weitere Datensätze aufzubereiten und sie über die Applikation verfügbar zu machen. Dies wären zum einen Taxonomien, die in unterschiedlichen DH-Disziplinen Verwendung findet. Hier bietet sich SKOS als primäres Input-Format an, da es weit verbreitet ist und entsprechende Transformationen von SKOS in das interne Graph-Format eine ganze Klasse von Datensätzen für die visuelle Erkundung in der Applikation zugänglich machen würde. Eine weitere Klasse potentieller Daten sind prosopographische Annotationen, durch deren Auswertung sogenannte Kookurenznetzwerke visualisiert werden können. Hierfür wurden schon Experimente mit Daten aus dem Schnitzler Tagebuch (8.500 Personen mit ca. 77.000 Nennungen) und aus der Zeitschrift ‚Die Fackel‘ (15.000 Personen mit über 123.000 Nennungen) durchgeführt.

¹ <http://clarin.eu/cmdi>

² <https://gephi.github.io/>

³ <http://www.graphviz.org/>

⁴ <https://github.com/mbostock/d3/>

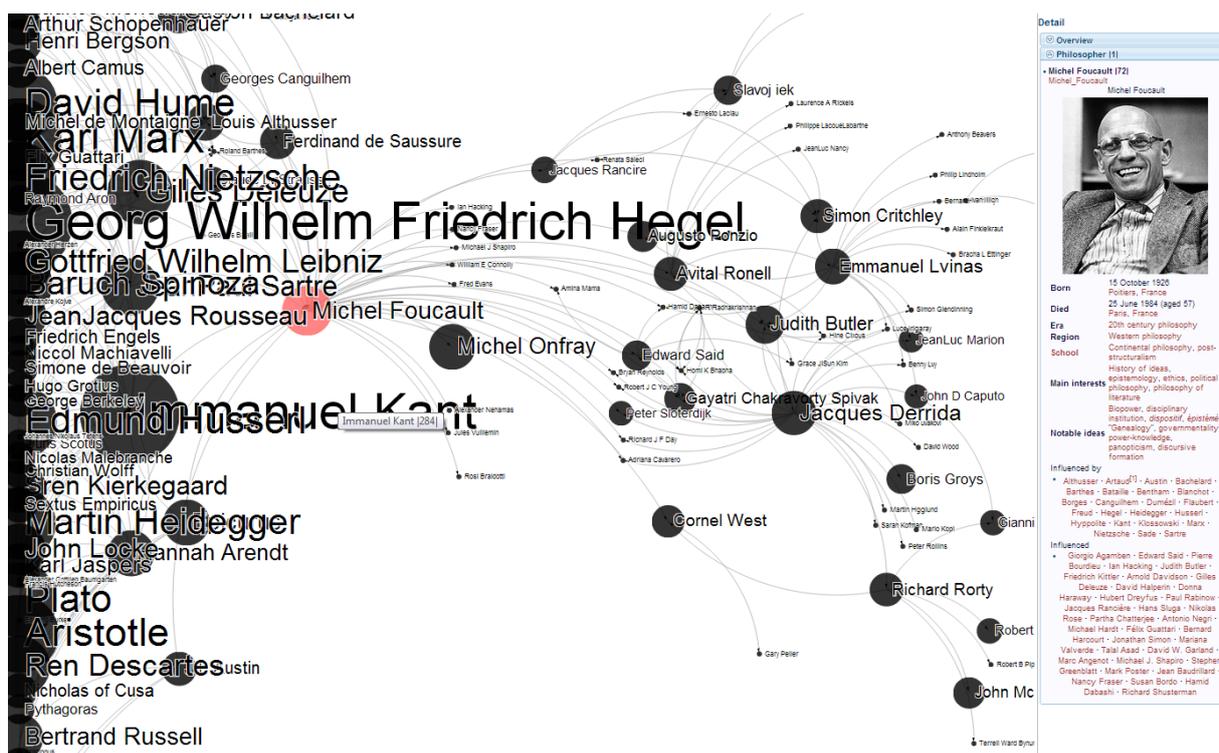


Abb 1 Screenshot der Visualisierung des "Philosophen-Influenz-Netzwerks". Es werden die VorgängerInnen und NachfolgerInnen von Michel Foucault dargestellt.

Nächste Schritte

Die Applikation ist bereits lauffähig und wird in ihrem ursprünglichen Kontext produktiv eingesetzt⁵, sie wird aber auch laufend weiterentwickelt. So ist eine Erweiterung geplant, um verbreitete standardisierte Graph-Formate (wie GraphML, GDF, GML) als Input und Output Formate zu unterstützen. Ebenso ist ein Refactoring des Codes notwendig, um die domänenspezifischen Aspekte von der generischen Applikation zu trennen und diese als ein sauberes wiederverwendbares konfigurierbares javascript-Modul anzubieten, das leicht in komplexere Applikationen eingebaut werden kann. Im Hinblick auf die eigentliche Visualisierung ist es wünschenswert, andere Darstellungsformen für die Knoten (momentan nur Kreise) anzubieten. Für tiefgreifende Analysen sind Graph-Operationen erforderlich, z.B. Vergleich von zwei Graphen und das Berechnen des gemeinsamen Subgraphen, Clustering u.ä.

Die Applikation wird als eigener Visualisierungsservice, der auch externe Daten annehmen, verarbeiten und visualisieren kann, als einer der Dienste des ACDH-ÖAW angeboten werden. Der Code wird open-source frei zur Verfügung gestellt.

Referenzen

- Broeder, D.; Kemps-Snijders, M.; Uytvanck, D. V.; Windhouwer, M.; Withers, P.; Wittenburg, P. & Zinn, C. (2010). A Data Category Registry- and Component-based Metadata Framework. In Calzolari, N.; Choukri, K. & others (Eds.). Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA).
- Đurčo, M. (2013). SMC4LRT - Semantic Mapping Component for Language Resources and Technology. Technical University, Vienna.

⁵ <http://clarin.oew.ac.at/smc-browser>

Ontologiestützte geisteswissenschaftliche Annotationen mit dem OWLnotator

Giuseppe Abrami

Alexander Mehler

Susanne Zeunert

Begriffe wie Annotationen, Relationen, Ontologien und Inferenz begegnen uns in allen Projekten, die sich mit der (geistes)wissenschaftlichen Erschließung von Korpora beschäftigen. Hierbei werden die Korpora mit den entsprechenden Annotationen des Anwendungsgebiets versehen, um auf dieser Grundlage Forschungsfragen zu beantworten. Annotationen sind ein wichtiges Mittel in der Analyse von Korpora; allerdings entwickeln die meisten Projekte ihre eigenen Strukturen und Formen der Annotations-Abbildung und -Verwaltung. Am Anfang der *Digital Humanities* wurden Annotations-Schemata teilweise fest codiert. Nunmehr werden vermehrt Beschreibungssprachen wie RDF Schema (RDFS) und die Web Ontologie Language (OWL)¹ eingesetzt. Da uns Ontologien flexible Annotationsmöglichkeiten bieten, jedoch die permanente Wartung und Anpassung von Software durch Informatiker auf lange Sicht keine effiziente Lösung ist, wurde in unserem interdisziplinären Projekt, gefördert durch LOEWE², zur inhaltlichen Erschließung der *Illustrationen zu Goethes Faust* der *OWLnotator*, ein ontologiebasiertes Annotationswerkzeug zur Erstellung und Analyse von Intra- und Intermedialen Relationen entwickelt. Dank der flexiblen Annotationsmöglichkeiten des *OWLnotators* können Geisteswissenschaftler durch das Erstellen von eigenen Ontologien sehr schnell und einfach ontologiestützt Annotationen im Einzel- oder Batch-Betrieb erstellen, ändern oder löschen.

Ein digitalisiertes Korpus von 2 500 Faustillustrationen bildet die Grundlage für die semantische Erschließung durch eine kunsthistorische Ontologie. Auf dieser Basis demonstrieren wir im Full-Paper die inter- und intramedialen Relationen zwischen dem Faust-Text und den dazugehörigen Bildern im *OWLnotator*. Das Korpus der *Illustrationen zu Goethes Faust* ist hierbei für diese Untersuchung in besonderer Weise geeignet, da einige Illustrationen Bildinhalte haben, welche im Text nicht erwähnt oder beschrieben wurden. Für eine hinreichend aussagekräftige inhaltliche Erschließung der Bildbestände ist es notwendig, die Bilder detailliert zu beschreiben. Dafür werden die Bilder *segmentiert* (cf. Abrami, Freiberg und Warner 2012) und detailliert annotiert. Zur Korpusverwaltung wird die ImageDB, ein Tool des *eHumanities Desktop* (Gleim, Mehler und Ernst 2012), verwendet, welche die Bildsegmentierung durchführt und als Annotation mittels des *OWLnotators* speichert. Der *eHumanities Desktop* ist eine plattformunabhängige, browserbasierte, flexible und skalierbare virtuelle Forschungsumgebung für Geisteswissenschaftler welche neben den genannten Tools weitere Werkzeuge zur Verwaltung, Analyse und Aufbereitung von Text- und Bild-Korpora wie auch von Lexika umfasst.

An den Korpora und den Annotationen können mehrere Forscher, in einer Arbeitsgruppe oder darüber hinaus, gleichzeitig arbeiten und je nach Forschungsschwerpunkt und Fragestellungen die annotierten Elemente entsprechend der gewünschten ontologischen Betrachtungsweise auswerten. Hierzu müssen die Forscher nur eine eigene Ontologie erstellen und diese im *OWLnotator* anwenden. Der *OWLnotator* kann jede syntaktisch gültige Ontologie nutzen und dank der in der Web Ontology Language spezifizierten Möglichkeiten der Klassen- und Relationsvererbung sowie der darin enthaltenen Inferenz-Methoden zur inhaltlichen Analyse des Korpus eingesetzt werden. Dank dieses ausdrucks mächtigen Werkzeugs sind wir künftig in der Lage, semantische Wissensnetzwerke sehr schnell und einfach aufzubauen und diese mit der Forschungsgemeinschaft zu teilen.

Unser Ziel ist es, durch den Einsatz von ontologischen Annotationen auf der Grundlage von OWL ein austauschbares Wissensnetzwerk zu generieren, welches unabhängig von der Software eingesetzt werden kann. Voraussetzung ist natürlich, dass die Software Ontologien lesen, interpretieren und verwalten kann, wie dies

¹<http://www.w3.org/TR/owl-features/>

²Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz, www.proloewe.de

durch den *OWLnotator* dynamisch und effizient geschieht. Die kunsthistorischen Ontologien unseres Projektes beinhalten die Annotationen der abgebildeten Personen auf den *Illustrationen zu Goethes Faust* sowie die Annotation ihrer Proxemik und Gesten. Durch atomare³ Annotationen schaffen wir die Grundlage zur Interpretation der annotierten Bildinhalte. Folgeprojekte oder ähnliche Fragestellungen können an den im Projekt erstellten Ontologien anknüpfen und darauf aufbauen sowie die Ontologien selbstverständlich erweitern. Die Ergebnisse sowie die Verknüpfung zwischen den Bildinhalten mit dem dazugehörigen Text sowie die weiter- und tiefergehende ontologische Annotation auf Text- und Bild-Ebene bieten für die Forschung, für die Lehre sowie für die Präsentation von Beständen und Korpora eine breite und stabile Grundlage und sind gleichzeitig jederzeit austauschbar und weiterverwendbar.

Durch unsere Arbeiten möchten wir Geisteswissenschaftlern die *Scheu* vor dem Einsatz digitaler Werkzeuge auch im Bereich komplexester Ontologien nehmen. Es geht darum, sehr große Korpora überhaupt erst auf der Basis dynamisch, im Wissenschaftsprozess wachsender Ontologien erschließbar zu machen. Mit dem *OWLnotator* bieten wir universell einsetzbares Annotationswerkzeug an welches Open Source zur Verfügung steht und allen Nutzern die Möglichkeit gibt, Annotationen ontologiebezogen durchzuführen, ohne über das *Wie* der Umsetzung nachdenken zu müssen. Insbesondere sollen Geisteswissenschaftler die Gelegenheit erhalten, nun über das *Was* des Annotationsinhaltes nachzudenken – dazu befähigt sie der *OWLnotator* in den Anwendungsgebieten der Geisteswissenschaft.

Literatur

- Abrami, Giuseppe, Michael Freiberg und Paul Warner (2012). „Managing and Annotating Historical Multimodal Corpora with the eHumanities Desktop - An outline of the current state of the LOEWE project Illustrations of Goethe's Faust “. In: *Proceedings of the Historical Corpora Conference, 6-9 December 2012, Frankfurt*.
- Gleim, Rüdiger, Alexander Mehler und Alexandra Ernst (2012). „SOA implementation of the eHumanities Desktop“. In: *Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Digital Humanities 2012, Hamburg, Germany*.

³Nicht mehr weiter teilbare

Sabine Seifert

Humboldt-Universität zu Berlin
Institut für deutsche Literatur
Nachwuchsgruppe „Berliner Intellektuelle 1800–1830“
sabine.seifert@hu-berlin.de

Poster Abstract**Gelehrsamkeit will verlinkt werden. Zur digitalen Erschließung von August Boeckhs Nachlass und Bibliothek**

Das Wirken August Boeckhs¹ (1785–1867), einer zentralen Figur im wissenschaftlichen Preußen des 19. Jahrhunderts, ist einer der Forschungsschwerpunkte der Nachwuchsgruppe „Berliner Intellektuelle 1800–1830“, geleitet von Dr. Anne Baillot an der Humboldt-Universität zu Berlin. Boeckhs bisher unedierte Handschriften bilden die Grundlage eines dreiteiligen Erschließungs- und Datenaufbereitungsvorhabens und werden auf folgende Weise zugänglich gemacht: 1) mittels einer digitalen Auswahl-edition, 2) durch die Rekonstruktion von Boeckhs Büchersammlung und 3) durch die Erschließung des handschriftlichen Nachlasses. Diese drei Bereiche sind eigenständige Forschungsunternehmen mit jeweils eigenen Ansprüchen und Forschungsfragen, die parallel ablaufen, sich aber durch die Verbindung auf digitaler Ebene gegenseitig befruchten und ergänzen.

Zu 1) Den Ausgangspunkt für das Poster soll die Edition ausgewählter Briefe und Berichte von und an Boeckh bilden, die im Rahmen der digitalen Edition „Briefe und Texte aus dem intellektuellen Berlin um 1800“² erfolgt. Diese führt nicht nur verschiedene Schriftsteller/-innen, Wissenschaftler und Intellektuelle zusammen, sondern auch verschiedene Textsorten und Themenschwerpunkte. Schon dadurch werden die edierten Handschriften Boeckhs in einem größeren, sie selbst übersteigenden Kontext verortet. Dieser Breite im Ansatz muss die zugrunde liegende Datenstruktur gerecht werden. Es wurden projekteigene Kodierungsrichtlinien³ nach den TEI P5 Guidelines⁴ entwickelt, die nur in Ausnahmefällen wie bei briefspezifischen Metadaten abgewandelt werden. Die Handschriften werden als Digitalisate zur Verfügung gestellt und die Transkriptionen in einer diplomatischen Umschrift sowie einer Lesefassung angeboten – ein Aspekt, der aufgrund der technischen Möglichkeiten relativ leicht zu realisieren ist, aber von kaum einer digitalen Edition tatsächlich umgesetzt wird. Die Auszeichnung von Personen, Orten, Werken und Organisationen und deren Erfassung in projektinternen Indizes ermöglichen eine umfassende Verknüpfung. Beides bildet die Grundlage für die Rekonstruktion (und geplante Visualisierung) der Netzwerke der Berliner Intellektuellen und für die Beantwortung der Forschungsfrage, wie sich diese Netzwerke entwickelt haben. Die Verwendung von Stan-

¹ Für neueste Forschungen siehe u.a.: Christiane Hackel, Sabine Seifert (Hrsg.), *August Boeckh. Philologie, Hermeneutik und Wissenschaftsorganisation*, Berlin 2013; Werther, Romy (Hrsg.), *Alexander von Humboldt. August Böckh. Briefwechsel*. Unter Mitarb. v. Eberhard Knobloch, Berlin 2011; Poiss, Thomas, „August Boeckh als Universitätspolitiker“, in: Anne Baillot (Hrsg.), *Netzwerke des Wissens*, Berlin 2011. S. 85–112.

² <http://tei.ibi.hu-berlin.de/berliner-intellektuelle/?de>

³ <http://tei.ibi.hu-berlin.de/berliner-intellektuelle/encoding-guidelines.pdf>

⁴ <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>

dards für webbasierte Textpräsentationen (XML/TEI, CC-BY-Lizenz, Normdaten, persistente URLs, ISO-Codes, etc.) bieten die Möglichkeit zu Kollaborationen mit anderen Projekten.

Zu 2) Die in den veröffentlichten Handschriften erwähnten Personen und Schriften finden sich häufig als Autoren und Werke in Boeckhs Büchersammlung wieder, die ca. 6000 Bände umfasste. Diese wird anhand einer von Boeckh selbst angefertigten und ebenfalls edierten Bücherliste virtuell rekonstruiert. Mit diesen Ergebnissen wird es möglich, Boeckhs Wissenshorizont und die Verbindungen zwischen Boeckhs eigenem wissenschaftlichen Arbeiten und dem seiner Fachkollegen und anderer zeitgenössischer Geisteswissenschaftler, ehemaliger Schüler und wissenschaftlicher Institutionen nachzuzeichnen. Darüber hinaus werden Boeckhs eigene Exemplare, aufbewahrt in der Universitätsbibliothek der Humboldt-Universität zu Berlin, auf Marginalien überprüft, die dann gegebenenfalls als Digitalisat zur Verfügung gestellt werden können.

Zu 3) Der bisher nicht an einer zentralen Stelle recherchierbare handschriftliche Nachlass Boeckhs soll, beginnend bei den verschiedenen Berliner Archiven und Bibliotheken, möglichst vollständig erschlossen werden. Für eine systematische und erstmalig institutionenübergreifende Darstellung dieser Daten wurde eine Plattform⁵ eingerichtet, die in ihrer technischen Struktur mit der Edition in Verbindung steht. Durch die Kooperation mit der Staatsbibliothek zu Berlin–PK konnten ca. 900 Kalliope-Einträge importiert werden.⁶ Da diese bibliothekarischen Einträge gerade in Bezug auf Boeckhs Korrespondenz teilweise eine sehr grobe Struktur aufweisen, wurden sie nach Autopsie und aufgrund von Forschungsergebnissen mit detaillierteren Metadaten zu jedem einzelnen Brief angereichert. Zusätzlich wurden, in Übereinstimmung mit der Edition, die in den Briefen genannten Personen, Werke etc. verzeichnet, um auch einen entitätenbezogenen Zugriff zu ermöglichen.

Nur durch die digitale Erfassung und Präsentation wird es möglich, dass sich diese drei Forschungsunternehmen in ihren wissenschaftlichen Ansätzen und Ergebnissen gegenseitig ergänzen und einen übersichtlichen und systematischen Zugriff für die Forschung bieten können. Durch die Indizes ist eine gleichzeitige Suche in den Handschriften der Edition, der Bücherliste und der mit ihr verbundenen bibliographischen Angaben und den Metadaten der Nachlassdokumente möglich. So können verschiedene Diskurse verfolgt und Tendenzen in der Philologie, in der Wissenschaftsorganisation in und über Preußen hinaus sichtbar gemacht werden. Die digitale Umgebung allgemein und konkret die Recherche und Nutzung derselben Datenbasis für gleichzeitig drei Unternehmen, die denselben Bezugspunkt – die Person Boeckh – haben, aber doch in sich eigenständig sind, können somit zur Forschung beitragen und Forschungsdaten zur Verfügung stellen, die von wissenschaftlichen Institutionen, Bibliotheken und Archiven genutzt werden können. Mit dem Poster möchte ich die digitale Edition in Verbindung mit der Rekonstruktion der Bibliothek und der Nachlasserschließung vorstellen und zeigen, welche Möglichkeiten digitale Methoden für die Forschung bieten und wie (Meta-)Daten als Schnittstelle zwischen Literaturwissenschaft, Bibliotheken und Archiven fungieren können.

⁵ <http://tei.ibi.hu-berlin.de/boeckh/>

⁶ <http://kalliope.staatsbibliothek-berlin.de/>

DFG-Projekt „Entwicklung eines MEI- und TEI-basierten Modells kontextueller Tiefenerschließung von Musikalienbeständen am Beispiel des Detmolder Hoftheaters im 19. Jahrhundert (1825–1875)“

Dr. Irmilind Capelle | Kristina Richts M.A., MA LIS

Die Kooperation von Bibliotheken und Wissenschaft erhält gegenwärtig vor dem Hintergrund des digitalen Wandels und der Entwicklung virtueller Forschungsumgebungen eine immer stärkere Bedeutung. Als Grundlage für eine solche Kooperation und die Zusammenführung der in unterschiedlichen Formaten vorliegenden Datenbestände ist die Entwicklung geeigneter Datenstandards unverzichtbar. Für den Bereich der Musikwissenschaft bringt der relativ junge Standard der Music Encoding Initiative (MEI) die für eine solche Zusammenführung notwendigen Anforderungen mit. Durch die Implementierung des Modells der Functional Requirements for Bibliographic Records (FRBR) im Jahr 2013 haben die Entwickler des Formats bereits einen entscheidenden Schritt in Richtung einer Zusammenführung mit in Bibliotheken vorliegenden Daten vollzogen. Das FRBR-Modell bildet dabei zum einen die Grundlage für das neue Katalogisierungsformat Resource Description and Access (RDA), zum anderen eignet es sich aber auch in besonderem Maße für die Beschreibung und Speicherung musikwissenschaftlicher Quellen. So sind erste Werkverzeichnisse dabei, die Vorteile des Modells zu nutzen – ein prominentes Beispiel ist der jüngst vom Danish Centre for Music Publication der Königlichen Bibliothek in Kopenhagen in digitaler Form veröffentlichte Carl Nielsen Works Catalogue (CNW). Im Rahmen des hier vorzustellenden, von der Deutschen Forschungsgemeinschaft (DFG) ab September 2014 über den Zeitraum von zunächst zwei Jahren geförderten Projekts steht die Entwicklung eines Modells zur kontextuellen Tiefenerschließung von Musikalienbeständen im Fokus. Vor dem Hintergrund der engen Kooperation von Bibliothek und Wissenschaft, die in Detmold ab Mitte 2015 auch räumlich und institutionell durch die Entstehung des neuen Zentrums „Wissenschaft | Bibliothek | Musik“ umgesetzt wird, beleuchtet das Projekt den Gegenstand sowohl von der wissenschaftlichen als auch von der bibliothekarischen Seite. So besteht das technische Ziel des Projekts darin, ein von anderen Bibliotheken mit vergleichbaren Beständen nachnutzbares Modell auf der Grundlage der XML-basierten Codierungsstandards der Music Encoding Initiative (MEI) sowie der Text Encoding Initiative (TEI) zu entwickeln, die beide sowohl eine bibliothekarische als auch eine wissenschaftliche Erfassung der Dokumente unterstützen und durch ihre Anbindung an internationale Datenstandards die Möglichkeit eines gezielten Mappings zu den Datenbeständen anderer Bibliotheken oder Forschungseinrichtungen mit sich bringen. Die bereits vorhandenen Vorteile speziell von MEI werden dabei gezielt im Hinblick auf ihre Anbindung an bibliothekarische Datenbestände weiterentwickelt und eine Anwendung erprobt.

Auf inhaltlicher Ebene sollen auf der Basis des entwickelten Modells die Vorteile einer kontextuellen Erschließung anhand des außergewöhnlich reichhaltig dokumentierten Musikalien- und Aktenbestands aus der Blütezeit des Detmolder Hoftheaters von 1825 bis 1875 demonstriert werden. Diese in der Lippischen Landesbibliothek Detmold erhaltenen musikalischen und archivalischen Quellen sind bislang entweder nur standardmäßig z. B. im Internationalen Quellenlexikon der Musik (RISM) (Musikalien) erfasst oder sogar lediglich durch

maschinenschriftliche Regesten (Theaterakten) sowie z. T. handschriftliche Zettelkataloge ausgewertet. Ergänzt werden sie durch Materialien aus dem Landesarchiv Detmold (Personalakten etc.) und dem Staatsarchiv Osnabrück (Theaterzettel).

In der ersten Projektphase geht es darum, die überlieferten musikalischen Quellen, die sowohl Partituren, Stimmen und Partien als auch Libretti und Rollenhefte umfassen, einerseits detailliert zu beschreiben (inkl. enthaltener Einlagen bzw. Striche sowie handschriftlicher Einträge zu Personen und Aufführungen) und andererseits die archivalischen Quellen im Volltext oder als Regesten zu erfassen. Im Rahmen der kontextuellen Tiefenerschließung sollen dann den erschlossenen Musikalien z. B. Informationen aus den Einnahme-Journalen oder den Regiebüchern des Theaters zugeordnet werden. So könnte in der Folge etwa ein mit Normdaten angereichertes Rollenverzeichnis aller mitwirkenden Schauspieler oder Sänger erstellt werden.

Um die Erkenntnisse, die aus dieser Erschließung der Daten entstehen, anschaulich zu visualisieren und möglichst offene Schnittstellen zur Weiternutzung zu bieten, werden die Projektergebnisse in einem Portal zusammengeführt, in dem Digitalisate der Materialien (in Auswahl) mit den XML-basierten Erschließungsdokumenten unter Rückgriff auf die in Detmold entwickelte Software Edirom Online verknüpft werden. Damit wird nicht nur für Forscher oder interessierte Laien eine Möglichkeit geschaffen, sich ein sehr viel präziseres Bild vom Wirken des Detmolder Hoftheaters in all seinen Facetten zu machen, sondern ein Repositorium geboten, das vielfältige Anknüpfungspunkte für weitere kulturwissenschaftliche Fragestellungen im Umkreis dieser wichtigen Institution des Hofes bietet.

ediarum – Eine digitale Arbeitsumgebung für Editionsprojekte

Stefan Dumont (dumont@bbaw.de), Martin Fechner (fechner@bbaw.de)

An der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) sind zahlreiche geisteswissenschaftliche Forschungsvorhaben unterschiedlichster Fachrichtungen angesiedelt. Die TELOTA-Arbeitsgruppe (»The Electronic Life of the Academy«) unterstützt diese Vorhaben in allen digitalen Belangen und entwickelt Softwarelösungen für die tägliche Forschungsarbeit der Wissenschaftler/-innen.

Die Erfahrung hat gezeigt, dass die Bereitschaft, TEI-Kodierung in Editionsprojekten zu verwenden, von der Benutzerfreundlichkeit der Eingabeoberfläche abhängt. Aus der Perspektive der Wissenschaftler/-innen erscheint es als ein Rückschritt, direkt im XML-Code zu arbeiten, wenn man vorher in Programmen wie MS Word gearbeitet hat. Eine neue Softwarelösung muss daher mindestens den gleichen Komfort bieten wie das zuvor benutzte Programm. Idealerweise würde sie sogar den gesamten Lebenszyklus einer Edition abdecken: von der ersten Phase der Transkription bis hin zur Publikation in Web und Druck.

TELOTA hat mit »ediarum« eine solche digitale Arbeitsumgebung entwickelt. Diese Lösung besteht aus mehreren Softwarekomponenten, die es den Wissenschaftler(inne)n erlauben, Transkriptionen von Manuskripten in TEI-XML anzufertigen, zu bearbeiten und zu veröffentlichen.

Als zentrale Softwarekomponente der neuen Arbeitsumgebung wird »oXygen XML Author« eingesetzt. Die Bearbeiter arbeiten in oXygen XML Author nicht in einer Codeansicht, sondern in der benutzerfreundlichen »Autorenansicht«, die über Cascading Stylesheets (CSS) gestaltet wird. Außerdem kann der Endanwender über eine eigene Werkzeugleiste per Knopfdruck Auszeichnungen vornehmen. So können z.B. in Manuskripten Streichungen markiert oder Sachanmerkungen eingegeben werden. Auch können Textstellen ausgezeichnet und gleichzeitig über eine komfortable Auswahlliste mit dem jeweiligen Eintrag eines zentralen Registers (Personen-, Ortsregister etc.) verknüpft werden. Der gesamte Text kann dadurch einfach und schnell mit TEI-konformen XML ausgezeichnet werden.

Die digitale Arbeitsumgebung nutzt die native XML-Datenbank »exist-db« als zentrales Repository für die XML-Dokumente. Die Datenbank ist auf einem Server installiert und online zugänglich. Dadurch können alle Projektmitarbeiter auf ein und denselben Datenbestand zugreifen und zusammenarbeiten.

Neben der eigentlichen Arbeitsumgebung in oXygen XML Author, wird für die Forschungsvorhaben auch jeweils eine Website auf Basis von eXist, XQuery und XSLT erstellt. In ihr kann von den Wissenschaftler(inne)n der aktuelle Datenbestand leicht durchblättert bzw. durchsucht werden. Die Website kann - je nach Bedarf - nur den Bearbeitern oder der gesamten Öffentlichkeit gemacht werden.

Als weitere Ausgabemöglichkeit wird mit Hilfe von ConTeXt eine Druckausgabe implementiert, die automatisch aus den aktuellen TEI-XML-Dokumenten ein PDF erstellt. Die Gestaltung und Formatierung kann - nach entsprechender Konfiguration - dabei gedruckten Bänden der jeweiligen Edition entsprechen. Jedem TEI-Element wird über eine Konfigurationsdatei eine entsprechende Formatierungsanweisung für den Druck übergeben. So können z.B. Text- und Sachapparat als Fußnoten dargestellt werden, die mit Hilfe von

Zeilennummerierung und Lemmata auf den Fließtext verweisen. Die Druckausgabe erstellt bei Bedarf auch das passende Register zu den jeweiligen Transkriptionen und löst Querverweise zwischen Texten auf.

Die Arbeitsumgebung wird seit 2012 von Wissenschaftler(inne)n verschiedener Forschungsvorhaben bei ihrer täglichen Arbeit benutzt. Nach ihrer Meinung befragt, waren sich die Nutzer darin einig, dass durch die neue Arbeitsumgebung die Editionsarbeit erleichtert und viel Zeit gespart wird. Auch die Möglichkeit, die Ergebnisse der Arbeit direkt in einer Webpräsentation oder Druckausgabe zu kontrollieren, wurde positiv gesehen. Sehr erleichtert äußerten sich die Mitarbeiter/-innen darüber, dass ihnen keine Arbeit im XML-Code selbst zugemutet wird, sondern alle Texte in einer grafischen und einfach zu bedienenden Programmoberfläche mit XML ausgezeichnet werden können.

Nach der erfolgreichen Pilotumsetzung im Akademievorhaben »Friedrich Schleiermacher in Berlin 1808-1834. Briefe, Vorlesungen, Tageskalender« wurde »ediarum« in zwei weiteren Akademievorhaben eingesetzt: »Commentaria in Aristotelem Graeca et Byzantina« und »Regesta Imperii - Friedrich III.« (letzteres in Kooperation mit der AdW Mainz) Für jedes Projekt wurden die TEI-XML-Schemata sowie die Funktionen an die verschiedenen Manuskripttypen und Forschungsanforderungen angepasst. Derzeit wird »ediarum« für die Historisch-kritische Edition der Schriften Jeremias Gotthelf zur Verfügung gestellt (in Kooperation mit der Universität Bern). Weitere Implementierungen befinden sich derzeit in Planung..

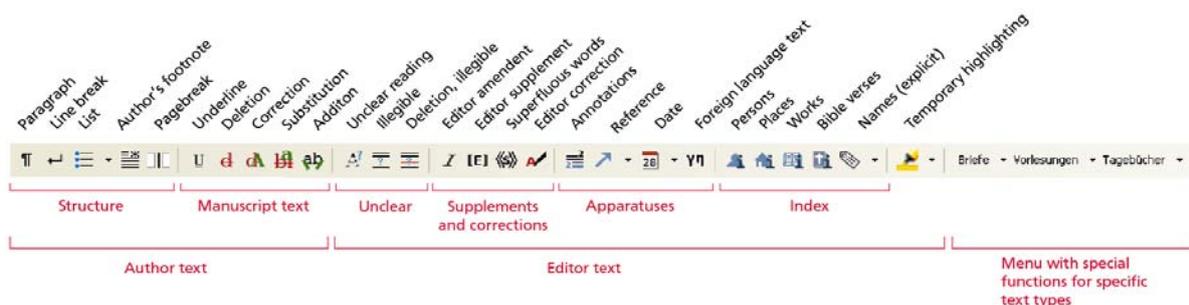
Weitere Informationen

- Projektwebsite: <http://www.bbaw.de/telota/software/ediarum>

Literatur

- Dumont, Stefan; Fechner, Martin: Digitale Arbeitsumgebung für das Editionsprojekt »Schleiermacher in Berlin 1808—1834« In: digiversity — Webmagazin für Informationstechnologie in den Geisteswissenschaften. URL: <http://digiversity.net/2012/digitale-arbeitsumgebung-fur-das-editionsprojekt-schleiermacher-in-berlin-1808-1834/>
- Burnard, Lou; Bauman, Syd (Hg.): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Charlottesville, Virginia, USA 2014. URL: <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>
- User Manual oXygen XML Author 14. URL: <http://www.oxygenxml.com/doc/ug-editor/>
- eXist Main Documentation. URL: <http://www.exist-db.org/exist/documentation.xml>
- ConTeXt Dokumentation. URL: http://wiki.contextgarden.net/Main_Page

Screenshots



Eigene Werkzeugleiste für die Schleiermacher-Edition in oXygen XML Author

Kombinierte Text- und Geo-Suche zum Durchsuchen einer Georeferenzierten Online-Bibliographie

Bastian Entrup¹, Vera Ermakova², Ines Schiller² und Henning Lobin²

¹Angewandte Sprachwissenschaft und Computerlinguistik

`bastian.entrup@germanistik.uni-giessen.de`

²Zentrum für Medien und Interaktivität

`{vera.ermakova|ines.schiller|henning.lobin}@zmi.uni-giessen.de`

Justus-Liebig Universität Gießen

Germany

1 Einleitung und Motivation

Das GeoBib Projekt entwickelt eine georeferenzierte Online-Bibliographie der frühen Holocaust- und Lagerliteratur zwischen 1933 und 1949 mit über 700 Werken und ca. 850 Autoren und Herausgebern. Anders als eine klassische Bibliographie werden auch handlungsrelevante Orte sowie biographische Daten inklusive einer schriftlichen Biographie zu Autoren und Herausgebern erfasst. Die bibliographischen Daten umfassen zusätzlich Informationen wie sie für ein Literaturlexikon nicht unüblich sind, z.B. Rezeptionen und Werksgeschichten. Die Kombination und Verlinkung dieser Entitäten, Personen, Werke und Orte, macht das Besondere und den Mehrwert der Online-Bibliographie aus.

2 Funktionen und Implementation

2.1 Funktionen und Implementation der Text-Suche

Um die entstehende Bibliographie und die dafür erstellten Texte (z.B. die Inhaltszusammenfassungen oder die Autorenbiographien) sowie die bibliographischen Daten (Autoren, Herausgeber, Verlag usw.) durchsuchbar zu machen wird Apache Solr ¹ und das Open Source Projekt *glp4lucene*² zur Verarbeitung von Suchanfragen und Erstellung des Indexes verwendet.

Die natürlich Variabilität und Ambiguität einer Sprache machen Verarbeitungsschritte aus dem Bereich des Natural Language Processings (NLP) notwendig. Im Bereich des Information Retrievals (IR) hat sich das Stemming als einfaches, regelbasiertes Verfahren zur Vereinheitlichung unterschiedlicher Wortformen auf einen gemeinsamen Stamm durchgesetzt (vgl. [3,5]). Aus linguistischer

¹ <https://lucene.apache.org/solr/>.

² Zu finden unter <https://sourceforge.net/projects/glpforlucene/>. Das Paket umfasst die Ergänzung von Synonymen, eine Lemmatisierungsfunktion sowie eine Termgewichtungsmethode, die auf der Wortart der Terme basiert.

Sicht ist Stemming jedoch nicht so erstrebenswert wie eine Lemmatisierung, die Reduktion von verschiedenen Wortformen auf ein gemeinsames Lemma, da beim Stemming die Ambiguität einer Sprache erhöht wird. Das hier genutzte Verfahren basiert auf dem MATE Tool [2] und verwendet das in [9] beschriebene deutsche Modell.

Basierend auf einem Lemma können Synonyme in GermaNet [4] nachgeschlagen und dem Suchindex hinzugefügt werden. Das Hinzufügen der Synonyme geschieht schon während der Indexierung der Daten³.

Die einfache textbasierte Suche durchsucht die wahrscheinlichsten Suchfelder nach einem Suchbegriff und nutzt dabei die Lemmatisierung des Indexes, um deklinierte oder konjugierte Formen zu finden. Zusätzlich sind im Index Synonyme vorhanden, so dass eine Suche nach *Gefängnis* auch Vorkommnisse von z.B. *Zuchthaus* findet. Die Suchergebnisse sind nach verschiedenen Personen-, Text- und Ortskategorien facettiert (s. Abb. 1).

Die Erweiterte-Suche liefert entweder Texte, Autoren/Herausgeber oder Orte als Ergebnis zurück. Wenn nach Texten gesucht wird, kann die Suche nach biographischen Daten der Autoren/Herausgeber (z.B. Name, Geburtsjahr oder Sterbeort), aber auch nach bibliographischen Daten (z.B. nach dem Verlag, dem Erscheinungsjahr oder -ort) gefiltert werden. Eine mögliche Suchanfrage wäre z.B. *Texte, deren Autoren weiblich sind*. Ähnliche Einschränkungen sind auch für Personensuchen möglich; z.B.: *Eine Autorin, die im Jahr 1939 einen oder mehrere Texte bei einem bestimmten Verlag veröffentlicht hat*.

2.2 Funktionen der Geo-Suche

Die Geo-Suche basiert auf einem Kartensatz Europas zur Zeit zwischen 1939 und 1945, der speziell für dieses Projekt aus verschiedenen Datensätzen kompiliert wurde (vgl. [6,7]). Für jedes Jahr wurde versucht, eine vollständige Karte mit den Grenzen Europas zu erstellen [8]. Die Jahre können über einen Slider unter der Karte ausgewählt werden. Auf der Karte dargestellte Datensätze können durch Klicken, Zoomen oder andere Werkzeuge ausgewählt werden.

Orte können in ein Suchfeld eingeben werden. Auf Grund der hohen Ambiguität von Toponymen wird dem User bei der Eingabe eines Ortsnamens eine Liste mit Vorschlägen angezeigt. Auf diese Weise gefundene Orte können dann mit einer Umkreissuche erweitert werden. Das ermöglicht zielgenaue regionale Recherchen, die für Heimatforscher und pädagogische Zwecke sinnvoll sind.

Ein Graph unterhalb der Karte (s. Abb 2) zeigt die Häufigkeit von Ereignissen (z.B. Anzahl von Handlungsorten) für jedes Jahr an. So können auf einen Blick Schwerpunkte ausgemacht werden. Ein Slider unter diesem Graph macht

³ Das ist bei dem relativ kleinen Datensatz vertretbar. Im Vergleich zu einer Verarbeitung während des Suchvorgangs, sorgt dies für eine geringere Auslastung und weniger Wartungsbedarf des resultierenden Systems. Das verwendete Software Paket erlaubt allerdings beide Möglichkeiten.

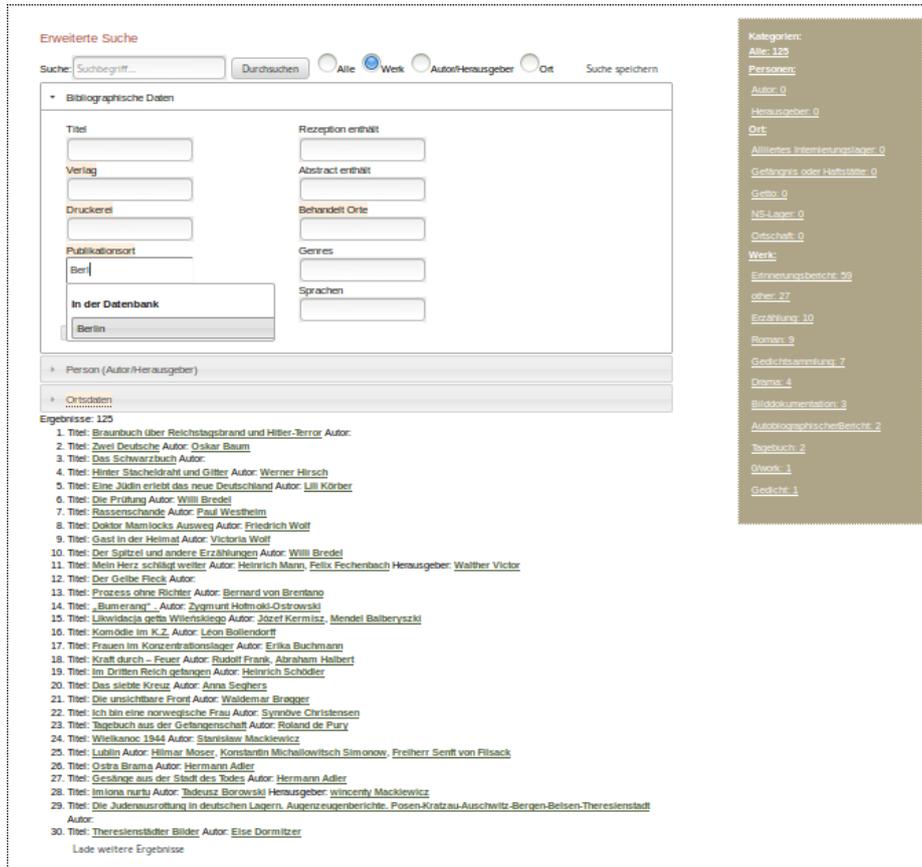


Abb. 1. Screenshot eines aktuellen Prototypen: Darstellung der Suchergebnisse und Vorschau der Eingabemaske.

es möglich sich nur Handlungsorte eines bestimmten Zeitraumes anzeigen zu lassen.

2.3 Verbindung von Text- und Geo-Suche

Die Verbindung der verschiedenen Entitäten in der Datenbank macht die Kombination der beiden Systeme möglich. Autoren/Herausgeber sind mit ihren Geburts- und Sterbeorten verbunden, außerdem mit den Orten in den von ihnen geschriebenen Texten. Werke sind mit ihren Erscheinungs- und Handlungsorten verbunden.

Die Beispiel-Suchanfragen können wie folgt ergänzt werden: *Texte, deren Autoren weiblich sind und in Berlin geboren wurden* und *Eine Autorin, die im Jahr*

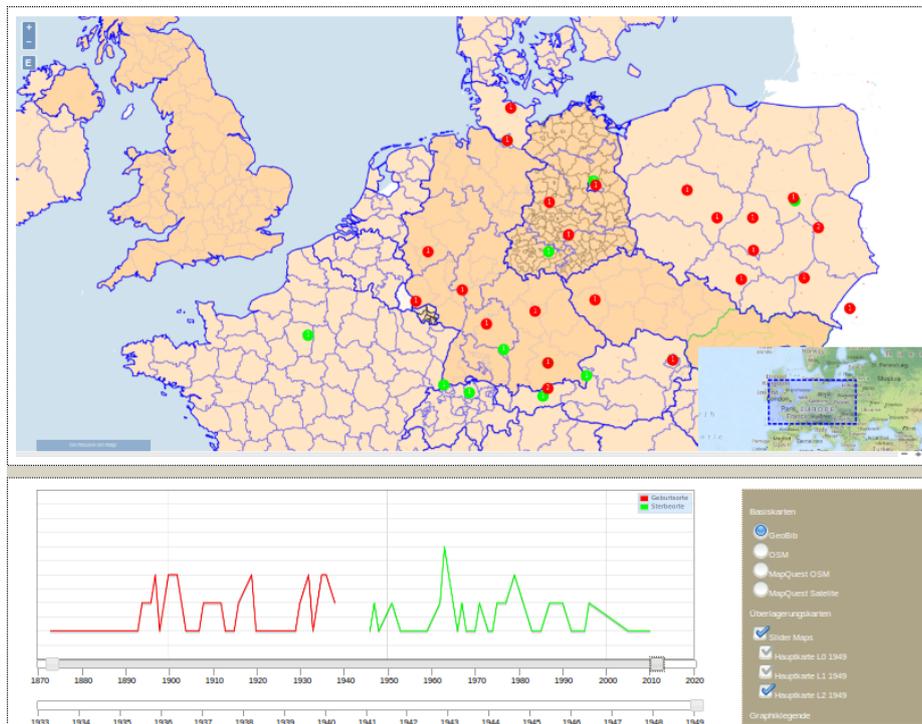


Abb. 2. Screenshot eines aktuellen Prototypen: Karte mit Angezeigten Geburts- (rot) und Sterbeorten (grün) basierend auf 125 Beispieltexen in den Grenzen von 1949.

1939 einen Text über Geschehnisse in Auschwitz bei einem bestimmten Verlag veröffentlicht hat. Auch Texte von Autoren, die in einer bestimmten Region geboren wurden oder Texte, die von einem bestimmten Lager handeln, sind so auffindbar.

Umgekehrt sind aber auch Orte auffindbar, die als Handlungsort zu bestimmten Zeiten eine Rolle spielen oder die Publikationsorte von bestimmten Werken sind. So lassen sich alle Orte finden, die in Werken eines bestimmten Autoren vorkommen.

3 Aussicht

Die Verbindung von Texten mit Geo-Daten ist nicht nur eine besondere Herausforderung an die Darstellung, die Organisation und die Suche nach Informationen, sondern bietet viele Möglichkeiten: Die Verteilung von Handlungsorten der Texte auf einer Karte bietet ein räumliches Verständnis eines Textes oder einer Sammlung von Texten. Besondere lokale Schwerpunkte können auf einen Blick erfasst werden.

Viele der im Projekt erfassten Texte gelten heute als vergessen. Sie werden nun das erste Mal systematisch durchsuchbar gemacht. Die Kombination von bibliographischen, biographischen, geographischen und inhaltlichen Daten ermöglicht einen völlig neuen (räumlichen) Zugang zu den Texten und den Ereignissen des Holocaust. So sind das Stellen und die Beantwortung neuer Forschungsfragen auf Grundlage einer breiten Textbasis und unter Berücksichtigung der geographischen Verteilung möglich.

Literatur

1. Binder, F., Entrup, B., Schiller, I., Lobin, H.: Uncertain about Uncertainty: Different Ways of Processing Fuzziness in Digital Humanities Data. In: Digital Humanities 2014, Book of Abstracts, pp. 97-100. Ecole Polytechnique Fédérale de Lausanne (EPFL) and The University of Lausanne (UNIL), Switzerland, 7-12 July 2014 (2014), <http://dharchive.org/paper/DH2014/Paper-874.xml>
2. Bohnet, B.: Very High Accuracy and Fast Dependency Parsing is Not a Contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 89-97. COLING '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
3. Braschler, M., Ripplinger, B.: How Effective is Stemming and Decomposing for German Text Retrieval? Information Retrieval 7(3-4), 291-316 (2004)
4. Hamp, B., Feldweg, H.: GermaNet - a Lexical-Semantic Net for German. In: Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. pp. 9-15 (1997)
5. Kraaij, W., Pohlmann, R.E.: Viewing Stemming as Recall Enhancement. In: In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 40-48 (1996)
6. Schaarschmidt, S.: Bestandserhebung zu verfügbaren digitalen geographischen Grundlagenkarten (2013), <http://geb.uni-giessen.de/geb/volltexte/2014/10572>
7. Schaarschmidt, S.: Bedarfsanalyse zu weiterem Kartenmaterial (2014), <http://geb.uni-giessen.de/geb/volltexte/2014/11102>
8. Schiller, I., Entrup, B., Binder, F., Schaarschmidt, S., Lobin, H.: Using a GIS for Search and Visualization of Literary Works in the Digital Humanities. In: gis.SCIENCE - Die Zeitschrift für Geoinformatik 4 (to appear) (2014)
9. Seeker, W., Kuhn, J.: Making Ellipses Explicit in Dependency Conversion for a German Treebank. In: LREC. pp. 3132-3139 (2012)

Digitalisierung der Universitätssammlungen der FAU Erlangen-Nürnberg

Auf dem Weg zum Semantic Web in Eigenregie

Martin Scholz und Udo Andraschke

(vorname.nachname@fau.de)

Die Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) besitzt über 20 Sammlungen aus den verschiedensten Fachbereichen. [1] Nicht minder verschieden gestalten sich Grad und Umfang ihrer Erfassung und Digitalisierung. Mit der Einrichtung einer Zentralkustodie im Jahr 2011, die die Aktivitäten und Ausrichtung der Sammlungen bündeln und sie als wichtige wissenschaftliche Infrastrukturen weiter ausbauen sollte, wurde auch das Ziel formuliert, die digitale Datenerfassung und Präsentation der Sammlungen voranzutreiben. Von zentraler Bedeutung sind dabei gemeinsame Erfassungsstandards und -formate sowie eine gemeinsame Software-Lösung und Webpräsenz.

Die Wahl der geeigneten Software-Infrastruktur fiel bewusst auf die Virtuelle Forschungsumgebung WissKI (wiss-ki.eu) [4], da sie

- a) unter einer Open Source-Lizenz verfügbar ist (GPL),
- b) konsequent auf offenen Standards und Formaten aufbaut und
- c) an der FAU mitentwickelt wird.

WissKI wird seit 2009 von der Arbeitsgruppe Digital Humanities des Departments für Informatik der FAU in Kooperation mit dem Germanischen Nationalmuseum in Nürnberg sowie dem Zoologischen Forschungsmuseum Alexander Koenig in Bonn als web-basiertes Content Management System für die Dokumentation von Kulturerbe im musealen und wissenschaftlichen Kontext entwickelt und befindet sich in den genannten Institutionen bereits im Einsatz. WissKI bietet den Nutzern gewohnte Eingabe- und Präsentationsschnittstellen, wie etwa feldbasierte Eingabemasken oder die Möglichkeit zum Freitext. Die Daten werden jedoch im Hintergrund nativ auf Basis von Semantic Web-Technologien (Ontologien, RDF [2]) erfasst. Dies ermöglicht auch technisch ungeschulten Nutzern das Einpflegen hoch vernetzter Datenbestände - sowohl lokal als auch global - und gleichzeitig das Erfassen der Bedeutung der Daten, um deren Interpretierbarkeit auf lange Zeit zu sichern. Dabei schreibt das System keine ontologischen Kategorien vor, sondern kann innerhalb des jeweiligen Anwendungsbereichs frei angepasst werden. WissKI ist nicht als zentraler Webdienst konzipiert, vielmehr kann die Software kostenlos heruntergeladen und auf einem Server als an die eigenen Bedürfnisse angepasste WissKI-Instanz eingesetzt werden.

Als fachübergreifende, verbindende Ontologie - eine sog. Referenzontologie - wurde der offene Standard CIDOC CRM [3] (ISO 21127) bzw. die OWL DL-Implementation Erlangen CRM (erlangen-crm.org) [5] gewählt, da das CIDOC CRM

- a) speziell auf die Dokumentation von Kulturerbe ausgerichtet ist und
- b) als international anerkannter Standard Sicherheit in Langzeitfragen gibt.

Die Referenzontologie garantiert zum einen ein Mindestmaß an fachübergreifender Interpretierbarkeit der Daten durch die Definition grundlegender Klassifikationsstrukturen und ermöglicht zum anderen die modulare Erweiterung um

fachspezifische Begrifflichkeiten.

Zur Umsetzung des angezeigten Vorhabens wurde das Pilotprojekt *WissKI@Sammlungen der FAU* [6] ins Leben gerufen. Partner sind neben der Zentralkustodie und der AG Digital Humanities drei ausgesuchte Universitätssammlungen: das Herbarium Erlangense, die Informatiksammlung Erlangen sowie die Schulgeschichtliche Sammlung. Die beteiligten Sammlungen spiegeln die oben genannte Heterogenität in hohem Maße wieder, so dass die unterschiedlichen Eigenarten und Bedürfnisse der Sammlungen der FAU weitgehend repräsentiert sind.

Das Pilotprojekt hat experimentellen Charakter. Es soll WissKI für den Einsatz in den Sammlungen erproben und einen Migrationspfad für die gesamten Universitätssammlungen entwickeln. Ein wichtiger Teilaspekt ist dabei der Transfer der Daten aus den bestehenden Datenbanksystemen in zuvor eingerichtete WissKI-Instanzen. Dennoch versteht sich das Vorhaben nicht als rein technikgetrieben, sondern sieht die Software als ein Instrument zum Ausbau der Sammlungen zu Forschungsinfrastrukturen.

Das Projekt startete im November 2013 ohne größere finanzielle Ausstattung. Treibende Kraft war und ist das Eigeninteresse der beteiligten Partner. Seit Mitte 2014 wird das Projekt durch eine studentische Hilfskraft unterstützt. Aufgrund der räumlichen Nähe aller Projektbeteiligten haben sich Workshops in regelmäßigen Abständen als Arbeitsmodus bewährt. Von großer Bedeutung ist dabei der gegenseitige Austausch, sowohl zwischen den Sammlungen untereinander als auch zwischen den Sammlungen und der Informatik, repräsentiert durch die AG Digital Humanities. Parallel dazu wurde eine WissKI-Instanz als Sandbox zum Üben eingerichtet und mit einem Forum und Wiki ausgestattet, so dass bspw. auch Tutorien erstellt und gemeinsam bearbeitet werden können. In der ersten Phase des Projekts (von Ende 2013 bis Mitte 2014) wurden monatliche Arbeitstreffen mit allen Projektteilnehmern anberaumt, um Themen zu behandeln, die alle Sammlungen angehen und um eine gemeinsame Wissensbasis zu schaffen. Nach ca. 6 Monaten wurden ergänzend Treffen zwischen der AG Digital Humanities (IT) und je einer Sammlung (Anwender) eingeführt, um den Spezifika der einzelnen Sammlungen besser Rechnung zu tragen.

Das Projekt wurde aufgrund der inhaltlichen Aufgaben in zwei Phasen unterteilt: Die bereits abgeschlossene Phase 1 beinhaltet alle Maßnahmen bis zum Transfer der Daten nach WissKI, die sich in fünf Schritte gliedern lassen:

1. Vertrautmachen mit der (Semantic) Web-Technologie und der WissKI-Infrastruktur
2. Identifikation von Gemeinsamkeiten und Unterschieden der bereits vorhandenen Daten und Datenbankschemata
3. Erstellen bzw. Erweiterung der Domänenontologie auf Basis des CIDOC CRM
4. Definition der Eingabemasken und Datenfelder und entsprechende Konfiguration der WissKI-Software
5. Iteration der Punkte 2-4

Phase 2 widmet sich dem Datentransfer und dem Aufbau eines gemeinsamen Portals in folgenden Schritten:

6. Definition der Abbildungsvorschriften zwischen bestehenden Datenbanken und Domänenontologie
7. Import von Testdaten aus den bestehenden Datenbanken

8. Test und Korrektur des vorgenommenen Imports
9. Iteration der Punkte 6-8
10. Einbinden bzw. Erstellen von Normdaten
11. Einbinden des Datenbestandes in ein gemeinsames Präsentationsportal

Alle Schritte wurden und werden begleitend in Form von Tutorien dokumentiert. Sie spiegeln Erfahrungen, Diskussionen und Best Practices wieder und bilden einen wichtigen Eckpfeiler für die Migration weiterer Sammlungen der FAU. Sie stellen in ihrer Gesamtheit eine stetig wachsende Gebrauchsanweisung für den Einsatz von WissKI auf der einen und Leitfaden zur semantische Modellierung von Sammlungen auf der anderen Seite dar. Die Tutorien sind öffentlich zugänglich und können auch Dritten außerhalb der FAU als Leitfäden dienen. [7]

Zum jetzigen Zeitpunkt (November 2014) ist Phase 1 abgeschlossen, Phase 2 befindet sich noch in der Umsetzung.

Im Vortrag soll daher auf Phase 2 nicht näher eingegangen werden. Vielmehr sollen Phase 1 analysiert und anhand von Beispielen einige der Hindernisse und Chancen des Vorgehens hervorgehoben werden:

1. Das Projektformat mit regelmäßigen Workshops und die aktive Einbindung von Sammlungsmitarbeitern setzt deren Bereitschaft voraus, sich eingehend mit aus Sammlungssicht meist fachfremden Methoden und Techniken auseinanderzusetzen. Im Bereich des Semantic Web handelt es sich zudem um ein relativ neues und dynamisches Gebiet der Informatik, das nicht mit jahrzehntelanger Erfahrung und entsprechend ausgereiften Werkzeugen aufwarten kann wie etwa relationale Datenbanksysteme. Naturgemäß bildet auch der Zeitaufwand (und damit indirekt die personellen Kapazitäten einer Sammlung) eine Hürde.
2. Im Gegenzug festigt das Format bei den Sammlungsmitarbeitern das Verständnis und die Akzeptanz für die eingesetzten Technologien und Methoden. Das Eigeninteresse der Sammlungen wird klar erkennbar, was in den Augen der Autoren die erfolgreiche Umsetzung des Vorhabens trotz begrenzter Mittel entscheidend begünstigt.
3. Nicht zuletzt bauen die Sammlungen über die beteiligten Mitarbeiter Kompetenz im Bereich IT und semantischer Modellierung auf. Die Sammlungen können sich im Idealfall untereinander austauschen und mithin gegenseitig helfen und unterstützen. Realistischerweise ist dies im Pilotprojekt bei einfacheren Fragestellungen der Bedienung und Modellierung gegeben.
4. Die ontologische Modellierung und der fächerübergreifende Charakter der Workshops führen zu einer vertieften Reflexion über die eigene Sammlung und den sammlungs- bzw. fächerübergreifenden Kontext. So wurde bspw. das Bewusstsein für fachspezifische Termini bei gleichzeitiger Notwendigkeit für gemeinsame Terminologien gestärkt. Das Auftreten teils unerwarteter Überschneidungen in den einzelnen Disziplinen ermöglichte u.a. auch eine genauere Definition sonst kaum weiter hinterfragter Begrifflichkeiten.
5. Das CIDOC CRM als Referenzontologie bietet hier einen guten Ausgangspunkt, um gemeinsame Strukturen und Prozesse trotz unterschiedlicher Fachbegriffe herauszuarbeiten und deren Bedeutung klar zu formulieren. Andererseits können die in der Referenzontologie vorgegebenen Strukturen zu zunächst eigenwilligen Ergebnissen führen. Die starke Betonung von Ereignissen im CIDOC CRM steht zum Beispiel im scheinbaren Widerspruch zur objektzentrierten Dokumentation vieler Sammlungen und erfordert teilweise ein Überdenken tradierter Muster und eine Korrektur institutionalisierter Denkgewohnheiten. Dies kann wiederum die oben genannte Reflexion anregen.

6. Durch die Analyse der Datenstrukturen werden darüber hinausgehende Gemeinsamkeiten sichtbar. Grundlegende Herausforderungen wie die Einbindung und Verwendung gemeinsamer Normdaten und die Sicherung der Datenqualität können benannt, diskutiert und einheitlich angegangen werden.

Nach Abschluss der ersten Projektphase kann somit resümiert werden, dass die Umsetzung des vorliegenden Vorhabens – der einheitlichen Digitalisierungen der Sammlungen der FAU in Eigenregie – zwar einen deutlichen Einsatz von den Sammlungen selbst einfordert, sich aber Mehrwerte ergeben haben, die für sie mittel- und langfristig von Vorteil sind.

Literatur & Internetseiten:

- [1] U. Andraschke und Marion Ruisinger, Die Sammlungen der Universität Erlangen-Nürnberg, 2007.
- [2] G. Antoniou, P. Groth und F. van Harmelen, A Semantic Web Primer, MIT Press, 2012.
- [3] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, Matthew Stiff (Hrsg.), Definition of the CIDOC Conceptual Reference Model, 2011.
- [4] M. Scholz und G. Goerz, WissKI: A Virtual Research Environment for Cultural Heritage. In Proceedings of ECAI. 2012, 1017-1018.
- [5] <http://erlangen-crm.org> (aufgerufen am 09.11.2014)
- [6] <http://wisski.cs.fau.de/sammlungen> (aufgerufen am 09.11.2014)
- [7] <http://wisski.cs.fau.de/sammlungen/tutorials> (aufgerufen am 09.11.2014)

Automatische Verfahren zur Bewertung der Relevanz von Dokumenten für geisteswissenschaftliche Forschungsfragen

André Blessing

Universität Stuttgart
andre.blessing@ims.uni-stuttgart.de

Melanie Dick

Universität Hildesheim
melaniedick@gmx.net

Ulrich Heid

Universität Hildesheim
heid@uni-hildesheim.de

Abstract

In vielen Projekten der Digital Humanities werden große Textmengen im Hinblick auf eine Forschungsfrage ausgewertet. Das interdisziplinäre Projekt *eldentity* (BMBF, FKZ. 01UG1234) widmet sich beispielsweise der Frage nach multiplen kollektiven Identitäten in internationalen Debatten um Krieg und Frieden seit dem Ende des Kalten Krieges. Damit sprachtechnologische Werkzeuge überhaupt auf das dort verwendete mehrsprachige Zeitungskorpus angewendet werden können, muss dieses zunächst von nicht für die Forschungsfrage relevanten Artikeln („off-topic-Artikel“) bereinigt werden. Nur so kann sichergestellt werden, dass nur Texte in die Auswertung einfließen, die Gegenstand der Forschungsfrage sind.

Viele Digital Humanities-Studien verwenden zur off-topic-Filterung lediglich Metadaten, wie zum Beispiel den Namen der Quelle, das Veröffentlichungsdatum oder den Autor. Für die Bereinigung des *eldentity*-Korpus genügen diese Informationen aber nicht: es muss zusätzlich eine inhaltliche Filterung vorgenommen werden. Kantner et al. (2011) erstellten dazu manuell Schlagwortlisten um relevante und nicht relevante Artikel zu identifizieren. Die Erstellung solcher Listen ist allerdings zeitaufwendig und in der Praxis stellten sich diese als nicht vollständig heraus. Die Aufgabe der off-topic-Filterung ist ähnlich wie sogenannte „Spam-Filter“ für e-Mail; allerdings kann ein Spam-Filter anhand großer Mengen von Daten trainiert werden, weil der Nutzer in der Regel alle Nachrichten manuell nach Relevanz bewertet. In Digital Humanities-Projekten ist die Anzahl der zu klassifizierenden Texte dafür zu groß; es braucht also Klassifikationsverfahren, die schon auf kleinen Mengen annotierter Texte gute Ergebnisse liefern.

In unserer Arbeit stellen wir einen neuen Ansatz vor; er geht aus von einer manuell annotierten Grundmenge von für die Forschungsfrage als relevant beziehungsweise irrelevant annotierten Artikeln. Ein Problem bei der Annotation ist die Auswahl der für das Training des Klassifikators nützlichen Artikel. Wenn zufällig ausgewählt wird, kann es sein, dass die Auswahl nicht repräsentativ für die zu klassifizierende Textmenge ist: wenn z.B. die Mehrheit aller Texte für die Forschungsfrage relevant ist, würden zu viele relevante und zu wenig irrelevante Texte annotiert. Hier kann durch die Nutzung von „Topic Modelling“ mit Latent Dirichlet Allocation (LDA; vgl. Blei et al. 2003) sichergestellt werden, dass eine besser nutzbare Auswahl getroffen wird.

Im ersten Schritt, der Feature-Extraktion, werden zunächst Merkmale aus Textdokumenten extrahiert. Diese extrahierten Merkmale werden in einem zweiten Schritt an einen Klassifikator übergeben, welcher die Artikel als relevant oder irrelevant kategorisiert (vgl. Abbildung 1).

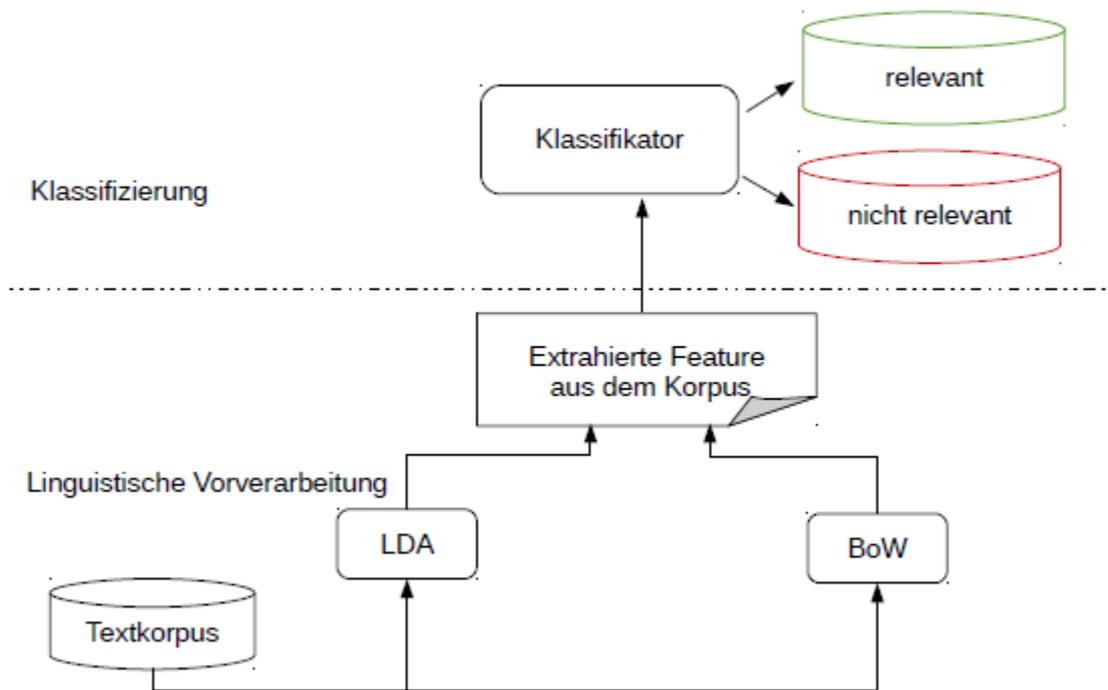
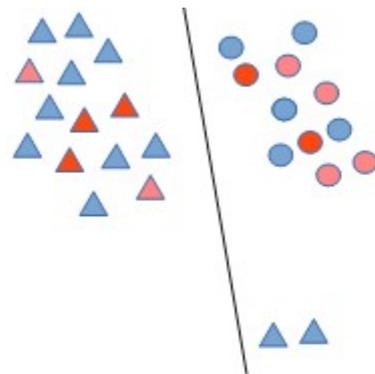


Abbildung 1: Zweistufiges Klassifikator-Modell

Für die Featureextraktion wird in unserem Experiment neben dem gängigen Bag-of-Words (BOW)-Modell, Topic Modelling durch LDA, ein generatives Wahrscheinlichkeitsmodell, eingesetzt und die Ergebnisse werden verglichen. Mit LDA können die Artikel entsprechend den vom System bestimmten „Topics“ vorsortiert werden. Ein „Topic“ soll mittels eines Wortclusters im abstrakten Sinne ein Themengebiet beschreiben. Der Klassifikator kann nun diese Topics explorieren (vgl. Abbildung 2) und anhand ihrer „Schlagwörter“ eine für den Forschungsgegenstand relevantere Auswahl kodieren. In unserem Forschungsprojekt konnten so sehr gut irrelevante Artikel zum Thema Sport, historische Konflikte, Buch- oder Filmkritik aufgespürt und als nicht relevant annotiert werden.



▲ relevant
● irrelevant

● ▲ Teil der Trainingsmenge

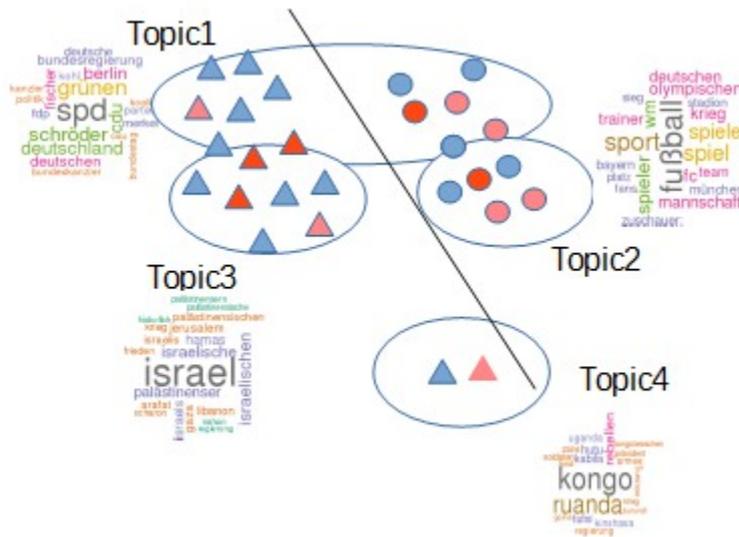


Abbildung 2: Exploration der Topics: oben: Zufallsauswahl; unten: LDA Topic-Modellierung

Im Rahmen der Vorverarbeitung wird zunächst eine Stopwortliste auf das Korpus angewendet. Häufig auftretende Wörter der geschlossenen Wortklassen werden damit als Merkmal ausgeschlossen. Als Baseline wird ein BoW-Modell verwendet, welches die Häufigkeit aller Wörter eines Dokuments als Hauptmerkmal für den Klassifikator aufbereitet. LDA hingegen lässt auf Wortebene für jedes Wort eine anteilige Zuordnung zu mehreren Topics zu. Die Information über die Verteilung über eine zuvor definierte Anzahl von Topics (beispielsweise 150) dient als Grundlage für den Klassifikator. LDA übergibt im Vergleich zum BoW-Modell Informationen an den Klassifikator, welche dieser wesentlich effizienter verarbeiten kann. Im zweiten Schritt folgt die Klassifikation in relevante und irrelevante Artikel (vgl. Abbildung 1).

Erste Ergebnisse haben gezeigt, dass das Topic Modelling für die Textklassifikation geeignet ist. Die manuell bewerteten Artikel (ca. 70 Dokumente – mehrfach bewertet) zeigten Accuracy-Werte von durchschnittlich 0,89. Erste Accuracy-Werte aus der automatischen Klassifikation (ca. 2.500 Dokumente – meist einfach

bewertet) mit LDA sowie BoW in der Vorverarbeitung, bewegen sich auch zwischen 0,8 und 0,9: die automatische Annotation liefert also die gleiche Qualität wie die menschlichen Annotatoren. Der Erfolg der Klassifizierung (Accuracy) beim LDA ist allerdings stark von der Anzahl der manuell ausgewählten Topics abhängig.

Weitere Variationen der beiden Verfahren wie beispielsweise eine Wortselektion mittels Part-of-Speech (POS)-Tagging sowie die Verwendung eines Stemmers, werden jeweils zusammen mit der linguistischen Vorverarbeitung eingesetzt: so kann deren Einfluss auf die Qualität der Ergebnisse des Klassifikators analysiert werden und die bestmögliche Merkmalsextraktion für die Entscheidung über die Artikelrelevanz kann bestimmt werden.

Literatur

David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003): *Latent dirichlet allocation*. IN: The Journal of Machine Learning Research 3, S. 993-1022.

Cathleen Kantner, Amelie Kutter, Andreas Hildebrandt, Mark Püttcher (2011): *How to get rid of the Noise in the Corpus: Cleaning Large Samples of Digital Newspaper Texts*, *International Relation Online Working Paper*. 2011/2, Juli 2011, Stuttgart: Universität Stuttgart.

eldentity (2014): *Multiple kollektive Identitäten in internationalen Debatten um Krieg und Frieden seit dem Ende des Kalten Krieges. Sprachtechnologische Werkzeuge und Methoden für die Analyse mehrsprachiger Textmengen in den Sozialwissenschaften (eldentity)*. URL: <http://www.uni-stuttgart.de/soz/ib/forschung/Forschungsprojekte/eldentity.html> Stand: 08.10.2014.

Vorschlag für ein POSTER:

ADHO Special Interest Group for Libraries and Digital Humanities

Special Interest Group (SIG) Organisatoren und Autoren:

Zoe Borovsky, UCLA Libraries Libraries, U.S.A.

Angela Courtney, Indiana University Libraries, U.S.A.

Isabel Galina, Universidad Nacional Autónoma de México

Stefanie Gehrke, Biblissima, France

Hege Stensrud Høsøien, National Library, Norway

Sarah Potvin, Texas A&M University Libraries, U.S.A.

Thomas Stäcker, Herzog August Library, Germany

Glen Worthey, Stanford University Libraries, U.S.A.

Das Poster hat zum Ziel, den Vorschlag einer Etablierung einer *ADHO Special Interest Group for Libraries and Digital Humanities* vorzustellen und im Rahmen der DHd 2015 zu diskutieren. Es wird die Gegenstände näher erläutern, die zu dem Antrag geführt haben, seine Genese als internationale Unternehmung darstellen und den Organisatoren dieser Gruppe die Möglichkeit geben, den Vorschlag im persönlichen Gespräch möglichen Unterstützern und an der Arbeit in der Gruppe Interessierten bekannt zu machen und zu erläutern. Konferenzteilnehmer, die nicht unmittelbar im Bibliotheksbereich arbeiten, soll das Poster verdeutlichen, von welcher Bedeutung DH heute in Bibliotheken ist.

Ziele

ADHO Libraries and DH SIG zielt darauf, die Zusammenarbeit und Kommunikation zwischen BibliothekarInnen und WissenschaftlerInnen zu fördern. Durch Einrichtung dieser SIG wird ADHO seinem Ziel gerecht, den Austausch zwischen den ADHO Organisationen und neuen DH Initiativen, die sich von bibliothekarischer Seite aus entwickeln, zu etablieren. Wir sind der Überzeugung, dass diese Verbindung zu einer sich intellektuell befruchtenden "doppelten Staatsbürgerschaft" führt, wo BibliothekarInnen und DH WissenschaftlerInnen gleichermaßen

in beiden Bereichen zu Hause sind. Durch die Förderung einer solchen „doppelten Staatsbürgerschaft“ werden Bibliotheken und BibliothekarInnen in die Lage versetzt, Möglichkeiten besser zu erkennen, wie sie sich in DH Projekte und Forschungsarbeiten einbringen können sowie insgesamt die Herausforderungen, vor denen sie stehen, besser zu bewältigen. Diese Herausforderungen schließen z.B. ein a) Finanzierungsmöglichkeiten zu ermitteln, Freistellungen und Schulungen zu ermöglichen, technische Infrastruktur bereitzustellen, um DH Projekte durchzuführen, b) die wechselnden Begriffe von „Dienstleistung“ und „Forschung“ mit Blick auf die meist kooperative Natur von DH Projekten zu hinterfragen und c) eine Kultur der digitalen Forschung in der Bibliothek zu etablieren.

Das Ziel der ADHO *Libraries and DH SIG* wird sein:

- Rat und Unterstützung anzubieten für die neu sich herausbildende Gruppe von BibliothekarInnen, die entweder eigene oder DH Projekte mit nicht der Bibliothek angehörig digitalen Geisteswissenschaftlern verfolgen,
- sich einzusetzen für Initiativen, die sowohl für Bibliotheken als auch DH von Interesse und von Vorteil sind (z.B. „Best Practices for TEI in Libraries“ und andere Richtlinien oder *best practice* Beispiele mit Bezug auf DH, die sich auf die Bibliothek beziehen)
- zu dokumentieren, wie sich BibliothekarInnen und Bibliotheken diesen Herausforderungen stellen
- Informationen zu liefern über verfügbare Ressourcen und Möglichkeiten (z.B. Schulungen, Drittmittel), die die Zusammenarbeit von verschiedenen, im Bereich der DH Forschenden, insbesondere in der Bibliothek, befördern,
- beispielhafte Projektergebnisse von BibliothekarInnen zu zeigen, die im Bereich der DH arbeiten,
- bibliothekarische Sichtweisen und Kompetenzen der gesamten DH community zu vermitteln.

Ein erstes Ziel der SIG wird sich darauf konzentrieren, Arbeitsbeziehungen zwischen internationalen Organisationen mit Bibliotheksbezug zu entwickeln, wie z.B. der ACRL DH Interest Group, der Digital Library Federation, der TEI in Libraries Special Interest Group, der

Society for American Archivists, der Association for Information Science and Technology oder der International Federation of Library Associations and Institutions.

Tätigkeiten

Mit Blick auf konkrete Aktivitäten würde die SIG sich einsetzen, um in Zusammenarbeit mit bibliothekarischen Organisationen Folgendes zu erreichen:

- Ermittlung und Nachweis von Bibliotheken, die DH Projekte durchführen und DH Organisationen, in denen Bibliotheken aktive Partner sind (z.B. das TEI Consortium SIG on Libraries)
- Konferenz-Sessions zu organisieren, einerseits für BibliothekarInnen auf DH Konferenzen, andererseits für andere DH Interessierte auf Tagungen, die sich in erster Linie an BibliothekarInnen richten (wie ALA, ACRL, ARLIS, DLF, Bibliothekartag, etc.)
- Workshops, Schulungen und Konferenz-Sessions zu organisieren, die dazu dienen, BibliothekarInnen stärker in die allgemeine DH Community zu integrieren und DH bezogene Bibliotheksprojekte vorzustellen.

Teilnehmen kann jeder, der Interesse an der Sache hat. Derzeit haben 130 Personen ihr Interesse bekundet, bei der SIG mitzuwirken. Wir denken jedoch, dass das potentielle Interesse weltweit weit höher liegt. Die Hoffnung besteht, dass durch die Posterpräsentation die SIG auch in der deutschsprachigen Community bekannter gemacht wird und neue Mitglieder geworben werden können.

Hinweise

Eine öffentliche Zotero Group zum Thema DH in libraries findet sich hier:
https://www.zotero.org/groups/adho_library_sig

Poster proposal DHd 2015

SMuFL-Browser und oXygen GlyphPicker Plugin

Werkzeuge zur Integration musikalischer Symbole in TEI

Alexander Erhard* Peter Stadler†

Die digitale Edition von musikalischen Texten und Texten über Musik bedarf an vielen Stellen der Darstellung musikalischer Zeichen und Symbole. Im Bereich 1D100–1D1FF des aktuellen Unicode-Standards sind zwar „Musical Symbols“¹ definiert, diese insgesamt 220 Zeichen decken aber nur einen Bruchteil des in der Praxis benötigten Repertoires ab. Eine breiter angelegte Systematik musikalischer Zeichen liegt in den Spezifikationen des *Standard Music Font Layout* (SMuFL)² vor, welche musikalischen Symbolen – ähnlich der *Medieval Unicode Font Initiative* (MUFI)³ im Bereich mittelalterlicher Zeichen – die Codepoints der Unicode Private Use Area zuordnen. Obwohl gegenwärtig kein Versuch unternommen wird, diese Zeichen in den offiziellen Unicode-Standard einzubringen, so stellt SMuFL doch für den Bereich musikalischer Symbole aufgrund seiner breiten Abdeckung einen de-facto-Standard dar.

Der Gebrauch von Unicode-Zeichen ist (neben dem Auszeichnen nach MEI oder MusicXML, dem Einbinden von Grafiken etc.) eine der von der TEI Music SIG in ihren Empfehlungen zu „TEI with Music Notation“⁴ diskutierten Möglichkeiten, musikalische Zeichen in TEI-Dokumenten zu repräsentieren. Einen

*Richard Strauss: Werke. Kritische Ausgabe, Universität München

†Carl-Maria-von-Weber-Gesamtausgabe, Universität Paderborn

¹Vgl. Perry Roland, *Proposal for Encoding Western Music Symbols in ISO/IEC 10646*, revised February 19, 1998, online verfügbar unter <https://archive.today/PzkaT>

²<http://www.smufl.org>

³<http://folk.uib.no/hnooh/mufi/>

⁴<http://www.tei-c.org/SIG/Music/twm/>

besonders geeigneten Ort hat die Verwendung von Musiksymbolen unseres Erachtens dort, wo einzelne musikalische Zeichen (oder kurze Sequenzen) losgelöst von einem größeren musikalischen Kontext in Worttext eingeflochten sind. Um das Finden und Einfügen dieser Symbole nach dem SMuFL-Standard in TEI Dokumenten zu vereinfachen, haben wir den Webservice „SMuFL-Browser“⁵ sowie das oXygen-Plugin „GlyphPicker“⁶ entwickelt.

Grundlage des Webservices sind Definitionen der mehr als 2000 Zeichen und Symbole im TEI-Format (mittels `<charDecl>` und `<char>`), die auf den SMuFL-Spezifikationen beruhen und als standardisierte Zielpunkte bei der Codierung von Musiksymbolen in TEI-Dokumenten (z. B. in der Form `<tei:g ref="http://mywebservice/smufl-browser/restQuarter"/>`) dienen können. Die Web-Oberfläche des SMuFL-Browsers erlaubt das bequeme Durchsuchen der Definitionen und stellt für jedes Musiksymbol neben Beispiel-Graphiken auch Code-Fragmente zum Einfügen in TEI-Dokumente bereit. Die Funktionalität orientiert sich an der ENRICH gBank application,⁷ geht aber durch die Bereitstellung einer REST-Schnittstelle für maschinelle Abfragen darüber hinaus. Via Content Negotiation werden Anfragen neben HTML auch in den Formaten TEI-XML oder JSON beantwortet, wodurch der Webservice auch als flexible Datengrundlage externer Tools dienen kann.

Die oXygen-Erweiterung „GlyphPicker“ versteht sich als Ergänzung zur regulären Zeichentabelle von oXygen und nutzt dafür den vorgenannten Webservice SMuFL-Browser. Sie unterstützt das Auffinden und Einfügen von Unicodefremden Zeichen, für die Definitionen in TEI mittels der Elemente `<char>` und `<charDecl>` vorliegen. Das Plugin bereitet entsprechende Definitionen zu Zeichentabellen in oXygen auf und stellt Mittel bereit, Verweise auf diese Definitionen (in Form von `<g>`-Elementen) in Dokumente einzufügen. Die Datenquellen der Zeichen-Definitionen sind im Plugin frei bestimmbar: Webservices wie der SMuFL-Browser werden ebenso unterstützt wie lokal abgelegte TEI-Dateien mit projektspezifischen Vorgaben. Das Plugin ist in der Texteditor- und Autor-Ansicht von oXygen nutzbar und kann sowohl selbständig als auch in Kombination mit einem Autor-Framework eingesetzt werden.

⁵<https://github.com/Edirom/SMuFL-Browser>

⁶<https://github.com/aerhard/glyphpicker>

⁷<http://www.manuscriptorium.com/apps/gbank/>

Dingler Dissemination

Highlights aus 6 Jahren »Digitalisierung des Polytechnischen Journals«

Marius Hug, M. A.; Martina Gödel, M. A.;
Timo Arndt, B.A.; Una Schäfer, B.A.

Einreichung zum Call for Posters

DHd-Tagung 2015

Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information
und Interpretation

7. November 2014

Abstract

Am 28. Februar 2015 endet die Laufzeit des von der DFG geförderten Projekts »Digitalisierung des Polytechnischen Journals« am Institut für Kulturwissenschaft der Humboldt Universität zu Berlin. Damit ist die DHd 2015 genau der richtige Zeitpunkt, um Bilanz zu ziehen. Mit unserem Posterbeitrag möchten wir die wichtigsten Ergebnisse aus der kooperativ angelegten Projektarbeit präsentieren. Highlights sind dabei: Nachhaltige Datenspeicherung und -präsentation, Ausführliche Tagging- und Projektdokumentation per ODD, Zurverfügungstellen der Daten über geeignete Schnittstellen zu Analyse Zwecken bspw. für Computerlinguisten, Aufbereitung der im TEI P5-Format vorliegenden Daten zur wissenschaftlichen Weiterverarbeitung (bspw. Umwandlung historischer Währungen, Visualisierung auf einer Timemap), Bearbeitung des sehr umfangreichen Bildmaterials mittels Image-Markup-Tool, sowie ein vollkommen neuartiger Transfer unserer Daten — aus der virtuellen Welt in die Welt der Objekte — als Grundlage für die Kooperation mit einem Museum.

1 Dingers Polytechnisches Journal (DPJ)

Das »Polytechnische Journal« wurde 1820 vom Augsburger Fabrikanten und Chemiker Johann Gottfried Dingler begründet. Dingler studierte wichtige Zeitschriften (die meisten davon aus England, Frankreich, später aber auch den USA), wählte relevante

Artikel aus, übersetzte und publizierte sie in seinem Journal. Mit einer Laufzeit von 111 Jahren ist diese Zeitschrift ein beispielloses, europaweites Archiv der Technik-, Wissens- und Kulturgeschichte. Besonders bemerkenswert ist die Aktualität der Publikation: So verging kaum Zeit zwischen Erstveröffentlichung der Artikel und Erscheinen der übersetzten Version im DPJ.

2 Das Digitalisierungsprojekt

Im von der DFG geförderten Projekt am Institut für Kulturwissenschaft der Humboldt-Universität zu Berlin wurde der komplette Bestand vom DPJ digitalisiert. Die Bilddigitalisierung wurde an der SLUB-Dresden durchgeführt. Für die Textdigitalisierung und Basisauszeichnung der über 200.000 Seiten war der Dienstleister Editura GmbH zuständig. Alle Bände sind per TEI-P5 kodiert. Das Journal ist online (CC by-nc-sa 3.0) unter www.polytechnischesjournal.de verfügbar.

Nachhaltigkeit, Clarin-D

Ein großes Problem für Digitalisierungsprojekte ist die nachhaltige Verfügbarmachung der Daten, Stichwort: Langzeitarchivierung. Für das Projekt Dingler-Online bedeutet die Zusammenarbeit mit dem BMBF geförderten Verbundprojekt **CLARIN-D** einerseits die Möglichkeit der Dissemination der Projektdaten, andererseits ist dadurch eine langfristige Sichtbarkeit des Projekts garantiert.

Durch die Kooperation mit der Berlin-Brandenburgischen Akademie der Wissenschaften, konkret dem DFG-Projekt **Deutsches Textarchiv** (DTA), profitiert Dingler-Online sehr direkt von deren technischem Know-how. Konkret zu nennen wären hier bspw. die orthographische Normalisierung und linguistische Analyse (POS, Tokenisierung, Lemmatisierung, ...) sowie die Nutzung der elaborierten Recherveschnittstelle.

Ausführliche Tagging- und Projektdokumentation per ODD

Alle editorischen Entscheidungen, die verwendeten Elemente und Attribute wurden in der ODD-Datei ausführlich beschrieben bzw. festgelegt. Für die eingesetzten Attribute wurden geschlossene Listen mit projektspezifischen Werten definiert und ihr Einsatz erklärt. Konkrete Quelltextbeispiele aus dem Projekt veranschaulichen das Vorgehen. Eine Transformation in das HTML-Format mittels des TEI-Tools OxGarage ermöglicht die Sichtbarmachung dieser Dokumentation im Look and Feel der TEI Guidelines selbst.

Das restriktiv formulierte Datenmodell und seine transparente Dokumentation unterstützen die möglichst schwellenarme automatisierte interne und externe Weiterverarbeitung. Die Dokumentation ist online unter <http://dingler.culture.hu-berlin.de/Schema/dingler.html> verfügbar.

Historische Daten

Ein wichtiges Thema – nicht zuletzt aufgrund der stets größer werdenden digital vorliegenden Datenmengen – ist die Visualisierung. Exemplarisch wurde in unserem Projekt ein solches Verfahren anhand von Patentdaten durchgeführt (s. Abb. 1). Diese eignen sich in besonderem Maße, da die Einheit Patentschrift sehr überschaubar ist, aber dennoch die für ein tief granuliertes TEI-Tagging benötigten Elemente enthält: Names, Dates, People, und Places (TEI P5 Guidelines, ch. 13). Der Workflow, mit Hilfe dessen die in Patentlisten vorliegenden Einträge in das zur Darstellung auf einer Timemap benötigte KML (Keyhole Markup Language) transformiert wurde, ist gut dokumentiert.

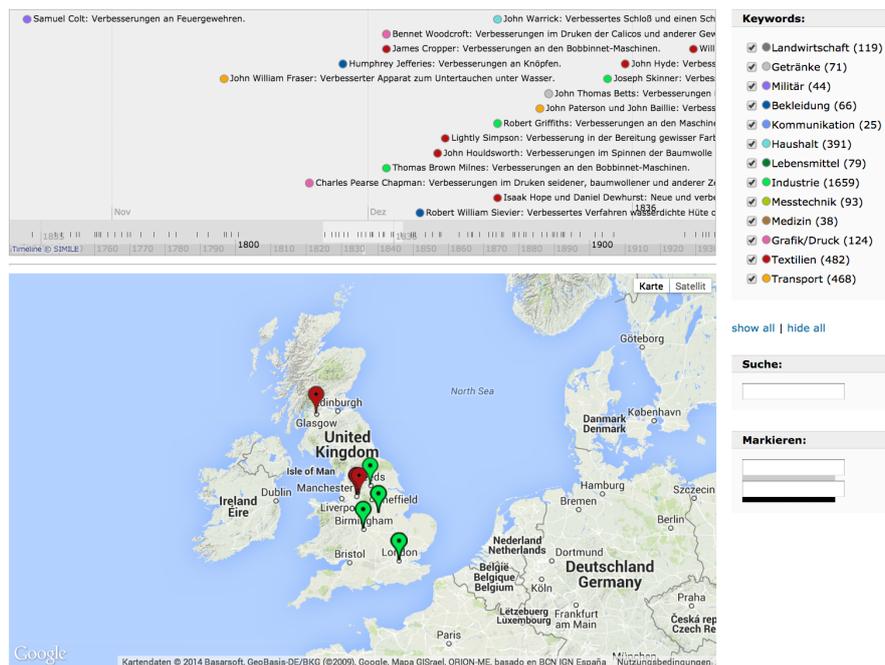


Abbildung 1: Visualisierung der TEI-Daten auf einer Timemap.

Historische Varianz stellt eine weitere Herausforderung für unsere Daten dar. Dies betrifft nicht nur die Texte, sondern auch Zahlen und Einheiten. Hier wurden die in den digitalisierten Daten vorkommenden Währungen und Einheiten gesammelt und über Listen entsprechende Umrechnungen zugewiesen, wobei sich bspw. für den Wiener Fuß folgendes Bild ergibt:

Image Markup Tool

Neben dem Textbestand (etwa 420 Mio. Zeichen) ist v. a. die große Menge an Bildern bzw. Zeichnungen hervorzuheben, wobei neben rund 3500 Falttafeln – eine Tafel enthält bis zu 114 Einzelfiguren – auch zahlreiche Figuren von Text umflossen gedruckt wurden. Um der Bedeutung des Bildmaterials im technischen Kontext gerecht zu werden, wurde

```
wienerfuss:  
  name: Wiener Fuß  
  unit: Wiener Fuß  
  wp: http://de.wikipedia.org/wiki/Fuß_(Einheit)  
  conversions:  
    zentimeter: x * 31.608  
    millimeter: x * 316.08
```

Abbildung 2: Syntax des projekteigenen Einheitenrechners.

hier bei der Erschließung besonders großer Aufwand betrieben.

Die Tafelwerke des »Polytechnischen Journals« wurden auf der Ebene der Einzelfiguren mit Koordinaten versehen. Der Mehrwert dieser Auszeichnungsstrategie besteht für den Nutzer zum einen in der konkreten Referentialisierung von Textpassage und Einzelfigur sowie individueller Anordnungs-, Betrachtungs- und Ausgabemöglichkeiten.

3 Museum

Hintergrund der von uns auf Grundlage unserer im Projekt aufbereiteten Daten angelegten Kooperation mit einem Museum – erste Prototypen sind gerade im Einsatz – ist folgende These: Weder ist Sammlung jenseits von Wissenschaft noch Forschung jenseits der Dinge möglich. Von einer Zusammenarbeit profitieren demnach beide Seiten. Ziel ist es, ausgewählte Objekte eines Museums multimedial erlebbar zu machen und damit den Besuchern ein neuartiges, interessantes Objekterlebnis zu ermöglichen. Andererseits erzeugt diese *Nachnutzung* unserer Daten eine längerfristige Sichtbarkeit der in den Projektdaten gespeicherten Informationen.

Mit Hilfe einer kostenlos zur Verfügung stehenden und speziell für das Projekt entwickelten App, werden dem Museumsbesucher zu ausgewählten Objekten vertiefende Informationen angeboten.

Denkbar sind hier bspw.:

- weiterführende Informationen zum Ausstellungsgegenstand und seiner Geschichte
- Informationen zu beteiligten Akteuren (Erfinder, Produzenten, Firmen...)
- veranschaulichende Bilder/Figuren
- verwandte Themenfelder, bspw. per Schlagwortwolke

Die App-Entwicklung wird bis zum Februar soweit sein, dass wir diese neuartige Nutzung unserer Daten tatsächlich hands-on vorführen können.¹

¹An dieser Stelle möchten wir darauf hinweisen, dass die Einreichung von Christian Kassung für einen Vortrag mit dem Titel »Making Things Chatter« eine weitere Technologie zur Verlinkung von Daten und Objekten präsentiert. Im Unterschied zu unserem Poster liegt der Fokus dort auf dem Museum.

4 Fazit

Zusammengefasst verfolgt unsere Posterpräsentation eine doppelte Strategie: 1) Das in der Community teilweise schon bekannte Projekt kann in verschiedenster Hinsicht Erfahrungen der letzten Jahre weitergeben und so für andere (kleinere) Digitalisierungsprojekte inspirierend sein. 2) Wir würden uns freuen, mit unserer neuartigen Idee einer Verknüpfung von Text- und Objektdaten (aus dem Museum) mit der Community in Diskussion zu treten und für das weitere Vorgehen von zu erwartenden Synergieeffekten zu profitieren.

5 Webressourcen

- CLARIN-D: <http://de.clarin.eu/de/>
- DinglerOnline: <http://www.polytechnischesjournal.de>
- Dingler ODD: <http://dingler.culture.hu-berlin.de/download>
- DTA: <http://www.deutschestextarchiv.de>
- Google Timemap: <https://code.google.com/p/timemap/>
- Image Markup Tool: http://tapor.uvic.ca/~mholmes/image_markup/
- KML: <https://developers.google.com/kml/>

Erweiterte Publikationen in den Geisteswissenschaften Zwischenergebnisse des DFG-Projektes Fu-PusH

Ben Kaden und Michael Kleineberg
Universitätsbibliothek der Humboldt-Universität zu Berlin, Deutschland

Das DFG-Projekt *Future Publications in den Humanities* (Fu-PusH), angesiedelt am Jacob-und-Wilhelm-Grimm-Zentrum der Humboldt-Universität zu Berlin, untersucht die Potentiale des digitalen Publizierens in den Geisteswissenschaften und erarbeitet szenarienbasiert Handlungsempfehlungen für akademische Infrastruktureinrichtungen wie insbesondere Universitätsbibliotheken und Rechenzentren, um den funktionalen Anforderungen unterschiedlicher geisteswissenschaftlicher Fachrichtungen gerecht zu werden.

Für Publikationsformen, die sich nicht mehr primär an der Druckkultur orientieren mit dem Versuch Printmedien etwa in Form von Monographien, Fachartikeln oder Sammelbandbeiträgen lediglich digital nachzubilden, sondern die genuinen Eigenschaften des Digitalen in den Mittelpunkt stellen, bietet sich die Bezeichnung *enhanced publications* bzw. „erweiterte Publikationen“ an. Solche Publikationsformen werden häufig als komplexe digitale Dokumente bzw. Dokumentensysteme charakterisiert, die sich unter anderem durch nicht-lineare Hypertextstrukturen, multimediale Zusatzmaterialien, integrierte Forschungsdaten, adaptive Darstellungsvarianten, dynamische Versionierung, kontextuelle Anreicherung sowie maschinenlesbare semantische Strukturierung auszeichnen. Ihre Vorteile liegen in einer engen Verknüpfbarkeit heterogener Elemente wie beispielsweise Digitalisate, Textkorpora, Datenbanken, Annotationen, Normdateien, Geoinformationen und narrativ-interpretativen Auseinandersetzungen mit diesen Objekten.

Auf diese Weise bieten erweiterte Publikationsformen die Möglichkeit nicht nur die Forschungsergebnisse, sondern auch die zu Grunde liegenden Forschungsdaten bzw. Forschungsprozesse in einem gemeinsamen Kontext zur Verfügung zu stellen, wobei die Grenzen zwischen Bearbeitungsraum, Kommunikationsraum und Veröffentlichungsraum sehr durchlässig werden.

Erweiterte Publikationen lassen sich demnach vor allem dadurch kennzeichnen, dass sie die in den Geisteswissenschaften etablierte Grundform der narrativen Auseinandersetzung mit einem Forschungsgegenstand an mindestens drei Stellen öffnen: Erstens kann ein direkter Bezug zu den Forschungsgrundlagen hergestellt werden, etwa durch eine Einbindung von bzw. Verlinkung zu digital vorliegenden Forschungsquellen wie Referenztexten, Abbildungen, Tondokumenten oder Filmsequenzen. Zweitens kann das narrative Element selbst über entsprechende semantische Tiefenauszeichnung durch Annotationen und Metadaten zu einem vielfältig vernetzbaren und maschinell prozessierbaren Datum werden. Drittens werden Interaktions- und Vernetzungsspuren solcher Dokumente wie beispielsweise Zitationen, Verlinkungen, Rezensionen, Verschlagwortungen oder Nutzungstatistiken darstell- und auswertbar.

Ob und inwieweit sich derartige Publikationskonzepte tatsächlich in der Praxis der Wissenschaften durchsetzen werden, hängt freilich vom Bedarf und auch der Bereitschaft der jeweiligen Fachgemeinschaften ab. Um auf diese Fragestellung einen substantiellen Zugriff zu erhalten werden im Fu-PusH-Projekt die Bedarfe, funktionale Anforderungen und Einstellungen systematisch in Interviews mit ExpertInnen aus dem Bereich der Geisteswissenschaften, aber auch mit Vertretern von Infrastruktureinrichtungen sowie Intermediären wie Verlagen und Anbietern alternativer Publikationsplattformen ermittelt.

Bei den zielgruppenorientierten Befragungen handelt es sich um qualitative und offene Leitfadeninterviews, die ein möglichst breites Spektrum an Perspektiven und thematischen Facetten abdecken sollen. Das Erhebungsinteresse schließt dabei neben technologischen Desiderata hinsichtlich digitaler Arbeits- und Publikationsumgebungen auch wissenschaftskulturelle, wissenschaftsstrukturelle sowie wissenschaftspolitische Anforderungen und Spielräume ausdrücklich ein.

In der Präsentation arbeiten wir zunächst den definitorischen Rahmen für erweiterte Publikationen heraus und spezifizieren funktionale Anforderungen an wissenschaftliche Veröffentlichungsverfahren. Im Anschluss setzen wir dies in Relation zu den Ergebnissen der Befragungen. Dabei differenzieren wir einen Ist-Zustand und einen auf einer Desiderats-Analyse basierenden Perspektiv-Zustand hinsichtlich der Publikationskulturen in verschiedenen geisteswissenschaftlichen Fachrichtungen. Auf diese Weise sollen aktuelle Transformationsprozesse in den Geisteswissenschaften sichtbar gemacht werden. Im Fokus stehen dabei insbesondere Einstellungs- und Handlungsmuster in Bezug auf:

- das wissenschaftliche Publizieren generell,
- die Erhebung, den Umgang sowie die Nachnutzung von Forschungsdaten,
- mögliche methodologischen Veränderungen unter dem Einfluss der Digital Humanities,
- das Publikationsverhalten insbesondere vor dem Hintergrund von Open Access,
- das Forschungsverhalten im Kontext von Open Science bzw. Open Scholarship,
- das Qualitätssicherungsverfahren des wissenschaftlichen Publizierens (Peer Review, etc.),
- die Dienstleistungen von Infrastruktureinrichtungen (z.B. Rechenzentren, Bibliotheken, Archive),
- die von Wissenschaftspolitik und Förderinstitutionen gesetzten Rahmenbedingungen,
- sowie mögliche Risiken im Zuge der digitalen Transformation.

Die Zwischenergebnisse des Fu-PusH-Projektes zeigen bereits sehr deutlich die Unterschiede im Forschungs- und Publikationsverhalten sowohl zwischen den Geisteswissenschaften und den so genannten MINT-Disziplinen als auch innerhalb des disziplinären Spektrums der Geisteswissenschaften selbst.

In diesem Zusammenhang soll die Frage verfolgt werden, inwieweit fachspezifische Publikationskulturen auch unterschiedliche technische und konzeptionelle Lösungen im Bereich der erweiterten Publikationen erfordern. Dies ist von besonderer Bedeutung, wenn man im Gegenzug die Herausforderung technischer Standardisierung zur Gewährleistung von Interoperabilität berücksichtigt. An dieser Stelle werden die Risiken deutlich, die generell von Technologien im Kontext der Digital Humanities ausgehen. Zum einen liegen bisher kaum Erfahrungswerte vor, mit denen sich eine tatsächliche Relevanzbewertung von Informationsinfrastrukturen bzw. Publikationsszenarien vornehmen lässt. Zum anderen besteht die Gefahr, dass neue technische Dispositive bestimmte Forschungs- und Erkenntnispraxen begünstigen und dafür andere weniger angemessen berücksichtigen.

Dies unterstreicht zusätzlich die Bedeutung der Modellierung komplexer Szenarien bevor Innovations-schritte angestoßen werden, da naturgemäß der Erfolg derartiger technischer Entwicklungen maßgeblich von der Passung mit dem tatsächlichen Bedarf und den Erwartungen – auch perspektivisch – der jeweiligen Zielgruppen abhängt. Insofern, und dies ist eine zentrale Erkenntnis auch dieses Projektes, müssen Schritte von Seiten der Infrastruktureinrichtungen, die die Forschungsrealität der Wissenschaftsgemeinschaften betreffen, im Dialog mit diesen erarbeitet werden.

Europeana Sounds – Ein Portal zu Europas klingendem Kulturerbe

Ute Sondergeld, Max Kaiser (Österreichische Nationalbibliothek, Wien)

Abstract

Die Massendigitalisierungsprojekte der vergangenen Jahre und die in diesem Zusammenhang entstandenen Portale und Repositorien haben zwar dazu beigetragen, den Zugang zu Primär- und Sekundärquellen des Kulturerbes zu erleichtern, unterliegen oftmals aber noch immer institutionellen oder regionalen Begrenzungen oder sind fokussiert auf bestimmte Dokumenttypen. Die Zusammenführung heterogener internationaler Datenbestände und verschiedener Informationstypen in einem zentralen Verweissystem sowie die Bereitstellung von Werkzeugen zu ihrer Bearbeitung und Weiterverarbeitung bietet die Chance, die Voraussetzung wissenschaftlichen Arbeitens – Sichtung, Auswahl und Kontextualisierung geeigneten Quellenmaterials zur Entwicklung von Forschungsfragen – zu stärken.

Das Projekt *Europeana Sounds* zielt darauf ab, die Datenbasis für den bisher weniger beachteten Themenbereich der Audioinhalte und der damit verwandten Dokumente innerhalb der digitalen Bibliothek *Europeana* zu stärken und so neben den bereits bestehenden Aggregatoren *APEX* (Archive), *EUScreen* (Fernsehen), *European Film Gateway* (Film) und *TEL* (Bibliotheken) die Infrastruktur für eine weitere Domäne der europäischen digitalen Bibliothek aufzubauen. Die Spannweite der im Projektrahmen zu referenzierenden Objekte reicht von Musik aller Sparten über Radiosendungen und Sprachaufnahmen bis hin zu Klanglandschaften, Natur- und Umweltgeräuschen. Der Einschluss verwandter Materialien wie Fotografien, Korrespondenzen, Textbücher, Musikdrucke und -handschriften trägt dazu bei, den Bestand audiobezogener Inhalte innerhalb der *Europeana* um mehr als das Doppelte auf insgesamt über eine Million Referenzen zu steigern und für Europa kulturell und historisch bedeutsame Objekte zentral zugänglich zu machen.

Grundlage der Datenaggregation bildet ein spezifisches, den Anforderungen von Audioobjekten entsprechendes *European Data Model Profile for Sound* sowie eigens entwickelte kontrollierte Vokabulare, die verschiedene Ebenen der referenzierten Objekte beschreiben. Basierend auf internationalen Normdaten tragen diese Vokabulare dazu bei, Metadatenqualität und das Retrieval multilingualer Daten, einer der großen Herausforderungen internationaler Datenbanken, sicher zu stellen.

Die Entwicklung von Tools zur Bearbeitung von Metadaten und Anwendungen zur Weiterverarbeitung von digitalen Objekten eröffnen Interaktionsmöglichkeiten mit dem Datenbestand. Zum Teil auf Anwendungen beruhend, die in anderen *Europeana*-Projekten entwickelt wurden, sollen die Werkzeuge zum Beispiel eine Korrektur und Transkription digitaler Objekte sowie ihre Klassifikation durch *social tagging* ermöglichen. Eine Kontextualisierung von Inhalten ist auf objektiver Ebene durch die Verlinkung zu ähnlichen Ressourcen oder Hintergrundinformationen sowie auf subjektiver Ebene durch persönliche Kommentare und Diskussionen vorgesehen. Zusammen mit der Möglichkeit einer individuellen Zusammenstellung von Objekten (Kuratierung) und der Einrichtung eines persönlichen Bereiches innerhalb des Portals wird eine Infrastruktur zur Verfügung gestellt, die sowohl dem allgemeinen Publikum, Experten wie auch Forschenden Möglichkeiten der Datenbearbeitung und -generierung bietet (Oomen & Aroyo, 2011; Chen, 2014).

Die Verbreiterung der Datenbasis durch die Zusammenführung verschiedener Bestände und die Bereitstellung einer Infrastruktur zu deren Weiterverarbeitung kann auf der einen Seite zu einer qualitativen Verbesserung des Informationssystems *Europeana* führen, eröffnet andererseits Möglichkeiten für die geisteswissenschaftliche Forschung und der Generierung neuen Wissens über das europäische Kulturerbe.

Das Projekt *Europeana Sounds* wird im Zeitraum von Februar 2014 bis Jänner 2017 von insgesamt 24 Institutionen aus 12 Ländern durchgeführt und von der Europäischen Kommission im Rahmen des ICT Policy Support Programme ko-finanziert. Die Österreichische Nationalbibliothek stellt im Rahmen des Projekts ihre wertvollsten Musikhandschriften von Komponisten des 17. bis 19. Jahrhunderts, die ihren Ruf als eine der bedeutendsten historischen Musiksammlungen weltweit begründen, zur Verfügung.

Literatur

Chen, C. (2014): Design for User Engagement on Europeana Channels. Master thesis, Delft University of Technology, Faculty of Industrial Design Engineering

Oomen, Johan & Aroyo, Lora: Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges.

Annotation und Analyse des literaturtheoretischen und -kritischen Diskurses in deutschsprachigen Poetiken (1770 bis 1960)

Die Poetik bildet als Wissensgebiet die theoretische Basis der Literatur- und Sprachwissenschaft von der Antike bis ins 20. Jahrhundert hinein. Als Poetik wird zugleich die Textsorte bezeichnet, die diese Theoriegrundlagen enthält und in der diese diskutiert und literaturkritisch geprüft werden. Im Zuge dieser diskursiven Verhandlung werden Verweise auf jeweils andere Poetik-Autoren und literarische Beispiele benutzt und teilweise kritisch bewertet. Die Analyse der quantitativen und qualitativen Aspekte dieser diskursiven Verweisungsstrukturen ist ein zentrales Ziel des Projekts ePoetics, einem BMBF-geförderten Kooperationsvorhaben der Universität Stuttgart und der Technischen Universität Darmstadt, und trägt zur Erforschung der Entwicklung grundlegender literaturtheoretischer Begriffe und Konzepte bei. Im Rahmen dieser Untersuchung wurden zwanzig deutschsprachige, für die Zeit von 1770 bis 1960 repräsentative Poetiken ausgewählt. Diese werden als TEI-konformes Corpus (inklusive der im Folgenden dargestellten Annotationsebenen) im Repositorium der virtuellen Forschungsumgebung TextGrid publiziert und für die weitere Erforschung nachnutzbar zur Verfügung gestellt. Darüber hinaus sollen auch im Projekt entwickelte Tools nachgenutzt werden können.

Um das Auftreten der zu untersuchenden vernetzten Verweisstrukturen in ihren unterschiedlichen Ausprägungen zu zeigen, aber auch das Vorgehen bei der Annotation und Analyse einzelner Begriffe und Konzepte zu erläutern, werden im geplanten Vortrag ‚Das Erhabene‘ und die ‚Metapher‘ als Beispiele hinzugezogen. Diese Begriffe eignen sich für Analysen von diskursiven Verweisungsstrukturen besonders gut, weil sie ihrer Herkunft nach aus der Ästhetik bzw. Rhetorik stammen und in der Poetik mit Rückverweis auf ihre Ursprünge und Urheber aufgeführt werden.

So geht das Konzept des Erhabenen als ambivalentes Gefühl der Überwältigung (bspw. bei der Betrachtung von Kunstwerken) zurück auf (Pseudo-) Longin, in dessen Nachfolge vor allem Kant und Burke den Begriff als ästhetische Kategorie definiert und umfassend diskutiert haben. Verweise auf diese beiden finden sich daher auch im Untersuchungscorpus wieder, allerdings mit entscheidenden Unterschieden. Denn das ästhetische Konzept des Erhabenen mit seinen zugehörigen Ersatz- und Ergänzungsbegriffen (Sublimes, Schreckliches usw.) erleidet im Untersuchungszeitraum einen Bedeutungsverlust, so wie die Ästhetik im Allgemeinen ihre poetologische Relevanz einbüßt, und wird entweder gar nicht mehr in seinem ursprünglichen Kontext behandelt, sondern nur noch im Zusammenhang mit erhabenem Stil innerhalb der *Genera dicendi*, oder erfährt ablehnende Beurteilung. Bei der Betrachtung des Begriffs und der

dazugehörigen Verweise muss insofern zwischen positiven und negativen Bewertungen unterschieden werden. Dies lässt sich durch das folgende Beispiel verdeutlichen: Bei Beyer (1882-84), Scherer (1888) und Wolff (1899) sind nicht nur zahlreiche Textstellen zu finden, in denen das Erhabene in seinem ursprünglichen Kontext behandelt wird, sondern auch jeweils Verweise auf Kant und/oder Burke als Urheber dieses Konzepts. Bei Beyer und Wolff findet eine produktive Auseinandersetzung mit dem Erhabenen statt, während Scherer den Begriff zwar in seiner ästhetischen wie poetologischen Auslegung nachverfolgt, ihn in seiner eigenen, empirisch ausgerichteten Poetik aber nicht reaktiviert, sondern rückblickend auf seinen Ursprung und seine Rezeption verwirft. D. h. Scherer nimmt zwar Bezug auf Kant, Burke und Autoren anderer literaturtheoretischer Werke, die über einen Diskurs des Erhabenen vernetzt sind, schließt sich diesem aber nicht an, sondern lehnt dessen Weiterführung ab.

Dieses erste Beispiel zeigt, dass die rein quantitative Identifikation, Annotation und Analyse relevanter Textstellen nicht ausreichen, um den Diskurs über das literaturtheoretische Konzept und dessen Entwicklung zu untersuchen. Eine komplexere Mehrebenen-Annotation ist erforderlich, die die Auszeichnung von Explikations- und Beschreibungskomponenten und Verweisstrukturen mit einer Bewertungsebene verbindet. Das für diese Anforderungen erstellte Annotationsschema umfasst daher konkrete Kategorien, die die Repräsentation des Begriffes und die Verweisungsstruktur im Text erfassen, und abstrakte Kategorien, die einerseits die Bewertungsebene abdecken, andererseits aber auch zur Überprüfung von Hypothesen dienen, die in einer vorhergehenden hermeneutischen Studie über das Corpus formuliert wurden (vgl. Sandra Richter: „A History of Poetics“). Im genannten Beispiel wäre die Hypothese, dass die Poetik im Ausgang aus dem 18. Jahrhundert noch expliziten Bezug auf die ästhetische Kategorie des Erhabenen nimmt und es unter Verweis auf Burke und/oder Kant diskutiert, während sich mit zunehmender Empirisierung der Poetik im Verlauf des 19. Jahrhunderts ein Bedeutungsverlust vollzieht und das Erhabene – wenn überhaupt – nur noch als stilistischer Aspekt thematisiert wird. Diesbezüglich wird ausgezeichnet, ob ein Bezug zur Ästhetik oder zur Stilistik besteht und inwiefern eine Bewertung erfolgt. So lässt sich nachvollziehen, dass eine derartige Entwicklung der Bedeutung des Konzepts Ergebnis eines wechselseitigen Diskurses ist.

Diese Diskursstruktur wird auch auf der Annotations-Ebene der konkreteren Repräsentation des Begriffes im Text erfasst. Dazu werden zusätzlich Verweisungen auf Personen bzw. Autoren und Werke ausgezeichnet (ebenfalls in Verbindung mit einer Bewertungsebene). Unterschieden wird dabei zwischen drei Textebenen: dem eigentlichen Poetikentext (Aussagen des Autors der jeweiligen Poetik), der Sekundärliteratur (Aussagen aus anderen literaturtheoretischen Texten, aber auch aus anderen Poetiken unseres Corpus) und der Primärliteratur (zur

Veranschaulichung herangezogene Beispiele aus literarischen Werken). Darüber hinaus wird bei gegebener Referenz auch zwischen den Verweisungsformen Zitat, expliziter und impliziter Paraphrase unterschieden. Vor allem letztere ist interessant, wenn sich nachweisen lässt, dass ein Autor einem anderen in seinen Ausführungen folgt, ohne dies anzugeben. Dadurch lassen sich über angegebene Verweisungsstrukturen hinweg auch unausgesprochene Übernahmen zurückverfolgen und ein diskursives Beziehungsgeflecht innerhalb des Corpus und darüber hinaus erfassen und sichtbar machen – auch auf der Ebene der literarischen Primärliteratur.

Dies lässt sich an einem zweiten Beispiel verdeutlichen. Bei der Definition der Metapher wird meist auf Aristoteles zurückgegriffen. Dieser versteht sie als „Übertragung“ zwischen einem eigentlichen und einem uneigentlichen Begriff und differenziert verschiedene Formen (vgl. Aristoteles: Poetik, Kap. 21). Diese grundlegende Definition lässt sich im Corpus „verfolgen“. Markant ist, dass einzelne Poetiken den Metaphern-Begriff als direkte Paraphrase von Aristoteles definieren und ihn gleichzeitig mit dessen teils literarischen Beispielen beschreibend darlegen. So finden sich in der Poetik von Borinski (1895) exakt dieselben Primärtext-Beispiele von Homer, die auch Aristoteles in seiner Poetik nennt. In einigen Poetiken lassen sich jedoch auch nur Ähnlichkeiten bei der Explikationsformulierung oder in der Unterscheidung verschiedener Metaphern-Formen bzw. Unterkategorien erkennen. Dies ist bspw. bei Beyer der Fall. Er definiert die Metapher als verkürzten Vergleich, bei dem der Vergleichspartikel wegfällt. Damit folgt er der Definition von Quintilian (neben Aristoteles die zweite grundlegende Begriffsbestimmung), ohne jedoch explizit die Quelle zu nennen. Darüber hinaus verweist er auf weitere Poetiken unseres Corpus, die den Begriff ebenfalls nach Quintilian bestimmen (Wackernagel 1873, Vischer 1846-57, Gottschall 1858). Das Beispiel zeigt, wie die Auslegung eines theoretischen Begriffs und dessen Entwicklung durch das Corpus hindurch mittels qualitativer Vergleiche und Analysen der (auch impliziten) Verweisstruktur nachweisbar ist.

Darüber hinaus wird die theoretische Definition eines griffigen Begriffs wie der Metapher häufig durch literarische Beispiele veranschaulicht, die sich für die Analyse auf der Ebene der Primärliteratur eignen. Zur Unterscheidung von Vergleich und Metapher verweist Beyer nicht allein auf Gottschall, sondern darüber hinaus auf ein bei diesem angeführtes Shakespeare-Zitat. An dieser Stelle verschränkt sich also die Analyse des Beziehungsgeflechts der Poetiken untereinander mit der Analyse der zitierten Primärliteratur. Literarische Beispiele werden in den Poetiken zur Veranschaulichung beschriebener Konzepte und für die (literatur-)kritische Stellungnahme im Hinblick auf die theoretischen Aspekte verwendet, sodass auch hier eine vergleichende Analyse möglich ist. Bestimmte Autoren und deren Werke tauchen in ähnlichen Zusammenhängen in einem großen Teil der Poetiken auf. Für die Metapher ist dies

Shakespeare. Clodius (1804), Gottschall und Dilthey (1887) nennen ihn übereinstimmend als einen der metaphernreichsten Dichter und damit als Vorbild für den richtigen Gebrauch von Metaphern. Ihm werden aber auch Negativbeispiele gegenübergestellt. Dies sind vor allem die antiken Autoren Sophokles und Aischylos, aber auch Goethe taucht in diesem unrühmlichen Zusammenhang immer wieder auf, was mit dem Unterschied von dessen epischem Stil zu Shakespeares dramatischem Stil begründet wird. Autorenbezogene Zuschreibungen wie diese lassen sich über das gesamte Corpus nachvollziehen und (auch diachron) vergleichend analysieren. Beispielsweise ist nach der jüngeren Poetik von Staiger nicht mehr Shakespeare der Prototyp des dramatischen Autors sondern Schiller, während Goethe diesem als Muster des lyrischen Dichters gegenübersteht. Durch solche Analysen ist es möglich, Prozesse der Kanonisierung und Ent-Kanonisierung einzelner Autoren und Werke nachzuvollziehen.

Sie helfen aber auch dabei, die Denkwelt einzelner Poetiken abzubilden. Bspw. erschien Wackernagels Werk zwar erst 1873 postum, es geht jedoch zurück auf eine akademische Vorlesungsreihe von 1836/7, was sich anhand der Auswahl der zitierten Primärliteratur einwandfrei nachvollziehen lässt. Ähnliches gilt für Autoren wie Staiger (1946) und Wehrli (1951), deren Schweizer Herkunft eine andere Textauswahl zumindest vermuten ließe. Während sich dies für Wehrli etwa im Hinblick auf Goethe bestätigen lässt, wird dieser für Staiger jedoch zum kanonischen Autor schlechthin.

Die computergestützte Auswertung all dieser Aspekte ermöglicht das Erkennen von Mustern und die Formulierung neuer Hypothesen bzw. Fragestellungen. Gleichzeitig bietet das Nebeneinander von abstrakter Interpretationsebene und konkreter Textebene Möglichkeiten des Abgleichs anhand der verschiedenen Kategorien. Das Annotationstool (UAM Corpus Tool) erlaubt eine Erweiterung des Schemas, sodass auch Aspekte, die im Verlauf der Untersuchung zu neuen Hypothesen führen, abgedeckt werden können. Auf der Basis der Auszeichnung der mit diesen Methoden selektierten Fundstellen nach dem beschriebenen Annotationsschema werden computergestützte Analysen und Visualisierungen durchgeführt. Auf diese Weise werden hermeneutische und algorithmische Verfahren im Sinne des ‚Algorithmic Criticism‘ verbunden. Dieses Vorgehen wird ausgeweitet auf weitere literaturtheoretische Kategorien auf verschiedenen Ebenen (z. B. Figur und Drama), sodass letztlich durch die kontextualisierende Aufbereitung, Vernetzung, Visualisierung und Analyse der enthaltenen Daten neue Erkenntnisse im Hinblick auf die Entwicklung der bedeutendsten literaturtheoretischen und -kritischen Diskurse und Konzepte in ihrem Zusammenhang ermöglicht werden.

Germania Sacra Online – Das Forschungsportal für kirchliche Personen und Institutionen bis 1810

Bärbel Kröger und Dr. Christian Popp, Germania Sacra

Wer im Netz nach wissenschaftlichen Informationen beispielsweise zu einem mittelalterlichen Kloster recherchiert, kann schon heute im Idealfall auf dem Forschungsportal der Germania Sacra ein ganzes Bündel von Informationen erhalten: Basisdaten zur Geschichte der Institution, kartographisch visualisierte Standortinformationen, Normdaten (GND, DBPedia, GeoNames), Links zu weiterführenden regionalen Online-Angeboten und bibliographische Informationen, Verknüpfungen zu dem in der Personendatenbank der Germania Sacra erfassten Klosterpersonal, die weitere fachübergreifende Verweise enthalten und damit den Weg zu neuen Erkenntnissen ermöglichen.

Die Germania Sacra hat in den vergangenen Jahren ein breites Portfolio digitaler Angebote zur Kirche des Alten Reiches erstellt. Hauptsäulen sind die digitalisierten Handbücher zur Geschichte kirchlicher Institutionen, die im Rahmen des Langzeitprojektes seit 1917 erarbeitet worden sind, ein umfangreiches digitales Personenregister zum kirchlichen Personal sowie die Datenbank zu Klöstern und Stiften des Alten Reiches. Alle digitalen Angebote der Germania Sacra sind work-in-progress: neue Bände werden nach 3 Jahren digital zur Verfügung gestellt, das Personenregister wird laufend erweitert (ca. 26.000 Einträge, Stand November 2014), die Klosterdatenbank befindet sich in der Aufbauphase (ca. 900 Einträge, Stand November 2014).

Die projektinterne Vernetzung der Daten wurde im Zuge der Integration in das Digitale Portal der Akademie der Wissenschaften zu Göttingen sichergestellt. Die hierfür verwendeten Technologien und Funktionalitäten werden im Rahmen des Vortrags präsentiert. Der Schwerpunkt des Vortrages wird jedoch auf der projektübergreifenden Vernetzung von Daten liegen. Ausführlich skizziert werden die hierfür bereits entwickelten Lösungsansätze sowie die zukunftsweisenden Kooperationen, die zu einer neuartigen digitalen Wissenslandschaft über die Kirche des Alten Reiches führen sollen.

Ein wichtiger Baustein für die Vernetzung ist die systematische Anreicherung der Datenbestände mit Normdaten. 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 Datenbanken, 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 Inhalte der Datenbanken auf der Basis von Linked Data angereichert und im RDF-Format ausgegeben, dabei 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.

Während bei der Verknüpfung von Daten zu kirchlichen Institutionen aufgrund eindeutiger Identifikatoren (z.B. Name, Orden, Standortinformation) vergleichsweise gut automatisierte Lösungsansätze zu finden sind, ist die automatisierte Vernetzung von personengeschichtlichen Datenbanken aus dem Bereich der Mittelalter- und Frühneuzeitforschung nach wie vor ein ungelöstes Problem im Bereich der Digital Humanities. Die Zuordnung von Informationen aus unterschiedlichen Datenbanken zu einer bestimmten Person ist schwierig. Häufig liegen nicht genügend Daten vor, die eine sichere Identifizierung ermöglichen (Geburts- und Sterbedatum, Herkunftsort, Ämter und Amtsdaten). Erschwerend kommen die zum Teil erheblich abweichenden Namensvarianten, Übersetzungs- und Transkriptionsfehler, latinisierte Formen und der spät einsetzende Gebrauch von Zweitnamen hinzu.

Daher entwickelt die Germania Sacra in Kooperation mit dem Deutschen Historischen Institut in Rom (DHI) und dem Repertorium Academicum Germanicum (RAG) eine projektübergreifende Personenrecherche. Dabei gilt es geeignete technische Lösungen zu finden. Hier können beispielsweise Algorithmen, die phonetische und orthographische Varianten auffindbar machen, oder die Verwendung von Thesauri zur Erkennung latinisierter Namensformen hilfreich sein. Diese Metasuche soll nicht nur als ein datenbankübergreifendes Recherchetool fungieren, sondern zugleich unter Verwendung von Technologien des Crowdsourcing interaktive Verknüpfungsmöglichkeiten für wissenschaftliche Nutzer bieten, die so ihre Identifikationsvorschläge einzelner Personendatensätze in die Datenbanken zurückmelden können.

Gerade die Verknüpfung von Personendaten aus unterschiedlichen Forschungsprojekten und unterschiedlichen Quellenbeständen lässt eine Generierung neuen Wissens erwarten, die in dieser Form nur durch den Einsatz digitaler Werkzeuge möglich ist.

Die nachhaltige Nutzung der Forschungsdaten wird durch die Integration von Germania Sacra Online in das Digitale Portal der Akademie der Wissenschaften zu Göttingen gewährleistet, die die hierfür erforderliche Infrastruktur zur Verfügung stellt.

Semantische Anreicherung von Bildnetzwerken mit HyperImage und Yenda – Das Hachiman Digital Handscrolls Projekt

Dr. Jens-Martin Loebel, Dipl.-Inf. Heinz-Günter Kuper

Abstract für ein Poster auf der DHd 2015 – Von Daten zu Erkenntnissen
23.-27. Februar 2015, Graz

Das Poster wird die Erweiterungen der virtuellen Forschungs- und Publikationsumgebung *HyperImage*¹ sowie die semantische Annotationsplattform [*Yenda*]² im Rahmen des Hachiman-Projekts vorstellen.

Das wesentliche Ziel des *Hachiman Digital Handscrolls* Projektes ist es, monumentale oder bewegliche Bildformate einer Forschungsgemeinschaft digital so vorzustellen, dass damit disziplinäre, sprachliche und regionalspezifische Grenzen aufgehoben werden. Inhaltlicher Gegenstand dieses Pilotprojekts ist ein Konvolut von sieben illuminierten japanischen Querrollen des 14. – 17. Jahrhunderts, die in leicht variierenden Versionen die Hachiman-Legende wiedergeben.

Die Materialität der Querrollen setzt einen physischen Kontakt und direkte Interaktion voraus: Sie müssen beim Ansehen mit beiden Händen entrollt werden und der genaue Bildabschnitt kann selbst bestimmt werden. Dieser Umstand und ihre Maße von bis zu 18 Metern pro Rolle müssen im Rahmen einer Untersuchung und Präsentation in einer Printpublikation oder statischen Datenbank zwangsläufig zu einem unbefriedigenden Ergebnis führen. Eine statische, auf Text- oder Bildabschnitte fixierte digitale Darstellung dieser Artefakte ist problematisch und in diesem Zusammenhang wenig zielführend.

Möglicherweise ist darin einer der Gründe zu suchen, weshalb bisher keine vergleichbare wissenschaftlich-digitale Aufbereitung solcher Querrollen mit gleichem Sujet erfolgt ist. Das Hachiman Digital Handscrolls Projekt möchte durch den Einsatz und gezielten Ausbau der virtuellen Forschungs- und Publikationsumgebung *HyperImage* und dessen Nachfolger Yenda zur semantischen Annotation genau diese Lücke schließen.

Durch Transkriptionen, Übersetzungen und visuelle sowie textuelle Annotationen, die gleichzeitig mit den dazugehörigen Textpassagen angezeigt werden können, wird den Betrachtern die inhaltliche Bedeutung vermittelt und der Text somit entmystifiziert und einem breiteren Publikum zugänglich gemacht.

HyperImage stellt dabei als technologische Basis Mittel und Werkzeuge bereit, die Rollen digital zu erschließen, zu annotieren, und – mittels des neuen Werkzeugs Yenda – semantische Verbindungen zwischen einzelnen Rollenabschnitten und Bild- und Textdetails zu visualisieren und diese mit Normdaten und Link-Open-Data-Beständen zu verknüpfen.

¹ siehe Website des Open-Source-Projekts unter <http://hyperimage.ws/>

² Die semantische Annotationsplattform [Yenda] wird ab Frühjahr 2015 als Open-Source zur Verfügung stehen und u. a. HyperImage als Web-basierte Arbeitsumgebung integrieren. Weitere Informationen finden sich unter <https://yenda.tools/>

HyperImage ist als Werkzeug zur Unterstützung des Bilddiskurses in den Digitalen Geisteswissenschaften seit vielen Jahren etabliert und wird in Forschung und Lehre an Forschungseinrichtungen in Deutschland und Europa eingesetzt. Mit HyperImage können beliebig viele Details innerhalb eines Bildes präzise markiert und beschrieben sowie Annotationen des Corpus untereinander verlinkt und über Indizes erschlossen werden können. Einfach gesagt: Was Hypertext für Text ist, ist HyperImage für Bilder (siehe Abb. 1 und Abb. 2).

Zwischenergebnisse wie endgültige Fassungen können jederzeit als hypermediale online- oder offline-Publikation erstellt werden. Verschiedene einzeln eingeführte und erprobte Verfahren und Datenrepositorien (u. a. das *prometheus Bildarchiv*)³ sind in HyperImage komfortabel zu einer einzeln oder kollaborativ nutzbaren Forschungs- und Publikationsumgebung zusammengeführt.

Ziel ist es, die digitale Darstellung von beweglichen Bild-und-Text-Formaten zu verbessern. Das innovative Open-Source-System HyperImage dient als zentrales Werkzeug, um sieben der japanischen Querrollen im Web zu präsentieren und somit historische Artefakte für zeitgenössische Sehgewohnheiten zu vermitteln und den Zugang zu historischem, literarischem und visuellem Wissen zu erleichtern.

Die Studie ist ein Kooperationsprojekt zwischen dem Institut für Kunstgeschichte Ostasiens, der Heidelberg Research Architecture (HRA) des Exzellenzclusters Asien und Europa, dem SFB 933 „Materiale Textkulturen“ und der Firma bitGilde IT Solutions UG.⁴ Die Digitalisate der Schriftrollen wurden von einer Reihe von Schreibern und Museen in Japan, den USA und Deutschland bereitgestellt (u. a. Hakozaki-gu, Kotozaki Hachiman-gu, Umi-Mori Art Museum, Yura Minato Jinja, Asian Art Museum San Francisco und Staatsbibliothek Berlin).

Die Berliner Firma bitGilde IT Solutions UG (eine Ausgründung der beiden Hauptentwickler von HyperImage) übernimmt die universitäts- und institutsübergreifende Koordinierung und nachhaltige Weiterentwicklung des Systems als Open Source. Mit der Ausgründung wird ein innovatives Konzept zur Verstetigung und Langzeitsicherung der Forschungsergebnisse aus Drittmittelprojekten verfolgt.

Durch das Projekt wird es erstmals möglich, diese Querrollen in digitaler Form einer breiten Öffentlichkeit über das Internet zugänglich und erfahrbar zu machen. Die Online-Veröffentlichung ist für Anfang 2015 geplant. Die Projektwebsite⁵ befindet sich derzeit im Aufbau.

Das Poster wird die zentralen Funktionen und Erweiterungen von HyperImage im Zusammenspiel mit der Annotationsplattform Yenda anhand der Ergebnisse des Hachiman-Projektes vorstellen.

³ <http://prometheus-bildarchiv.de>

⁴ siehe <http://bitgilde.de>

⁵ <http://www.zo.uni-heidelberg.de/iko/hdh/>

Abbildungen

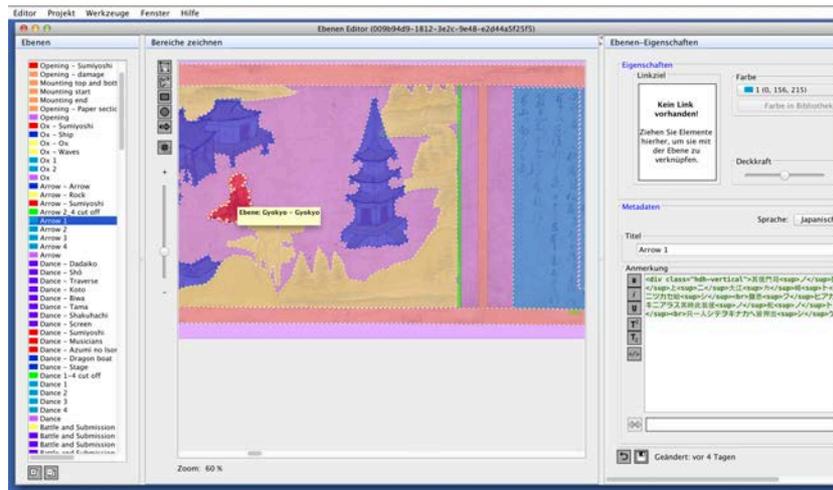


Abb. 1: Visuelle Annotation und Transliteration der Rollen mit HyperImage.

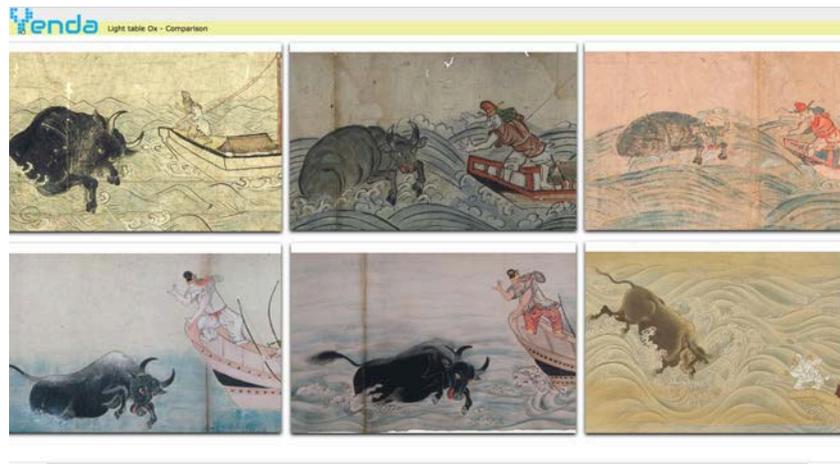
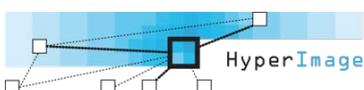


Abb. 2: Living Scrolls-Technologie von Yenda. Gleichzeitige Visualisierung ein und desselben Motivs von sechs verschiedenen Rollen. Jede Rolle ist einzeln in der Webpublikation im Browser navigierbar.

Weiterführende Literatur

Loebel, J.-M., Kuper, H.-G.; et al.: Hachiman Digital Handscrolls – Semantische Anreicherung mit HyperImage und Yenda. In: Bienert, A.; Hemsley, J.; Santos, P. (Hrsg.): *EVA Berlin 2014 – Elektronische Medien & Kunst, Kultur und Historie*. Berlin: Konferenzband, ISBN: 978-3-88609-755-5, 2014, S. 262-267.

Kuper, H.-G.; Loebel, J.-M.: HyperImage: Of Layers, Labels and Links. In: *Proceedings of RENEW – the 5th edition of the International Conference on the Histories of Media Art, Science and Technology*, Riga, 2014.



UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386



Humanities Data Centre – grundlegende Überlegungen in der Designphase eines geisteswissenschaftlichen Forschungsdatenzentrums

Stefan Buddenbohm¹, Claudia Engelhardt², Ulrike Wuttke³

¹Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften, ²Staats- und Universitätsbibliothek Göttingen, ³Akademie der Wissenschaften zu Göttingen
buddenbohm@mmg.mpg.de, claudia.engelhardt@sub.uni-goettingen.de, uwuttke@gwdg.de

Schlagwörter: Forschungsdatenzentrum, Langzeitarchivierung, Forschungsdatenmanagement

Zusammenfassung: Forschungsdaten sind sowohl Ergebnis von Forschung als auch Grundlage für neue Forschungsfragen. Die zunehmende Nutzung digitaler Ressourcen und Methoden in der Forschung widerspiegelt sich sowohl im wachsenden Umfang als auch in der zunehmenden Komplexität von digitalen Forschungsdaten, sowohl in den Geisteswissenschaften wie auch in anderen Disziplinen. Aus verschiedenen Gründen ist die Erhaltung dieser Forschungsdaten notwendig: Dokumentationszwecke beispielsweise für Förderer oder aufgrund rechtlicher Bestimmungen, Nachvollziehbarkeit und Reproduzierbarkeit von Forschungsergebnissen, aber vor allem auch die Möglichkeit der Nachnutzung für neue Forschungsvorhaben. Die Herausforderungen hinsichtlich des Forschungsdatenmanagement und der Langzeitarchivierung können jedoch nur mit einem umfassenden Verständnis ihrer Entstehungs- und Nutzungsbedingungen gemeistert werden. Da diese von Infrastrukturanbietern nur im engen Austausch mit den Fachdisziplinen eruiert werden können, scheinen disziplinspezifische Forschungsdatenzentren am besten geeignet, die damit verbundenen Aufgaben zu übernehmen.

Während der Designphase des Humanities Data Centres (HDC, 2014-2016) werden daher im Dialog mit der Wissenschaft und Infrastruktureinrichtungen die Grundlagen für den Aufbau eines Forschungsdatenzentrums für die Geisteswissenschaften geschaffen. Das Projektkonsortium besteht neben geisteswissenschaftlichen Forschungseinrichtungen aus Rechenzentren und einer Universitätsbibliothek.

Grundsätzlich lässt sich die Langzeitarchivierung von Forschungsdaten entlang von drei aufeinander aufbauenden Ebenen strukturieren:

- Bitstream Preservation: Der physische Erhalt des gespeicherten Datenobjekts (Bitstream) auf einem entsprechenden Speichermedium,
- Technische Nachnutzbarkeit: Sicherstellung der Zugänglichkeit der Forschungsdaten auch bei veränderten technischen Bedingungen,
- Intellektuelle Nachnutzbarkeit: Sicherstellung der vollständigen Nutzbarkeit und Interpretierbarkeit des intellektuellen Gehalts der Forschungsdaten, beispielsweise durch Metadaten und die Dokumentation von Kontextinformationen, die das ursprüngliche Forschungsszenario nachvollziehbar machen.

Darüber hinaus hängt die Nachhaltigkeit von Forschungsdaten stark von einem stabilen, organisatorischen Rahmen ab, innerhalb dessen die entsprechenden Umgebungen und Werkzeuge bereitgestellt werden können. Nicht zuletzt ist aber der beständige Austausch mit den wissenschaftlichen Nutzern von großer Bedeutung, um mit dem Angebot (dem Forschungsdatenzentrum) den Anforderungen der Wissenschaft zu entsprechen beziehungsweise dieses neuen Entwicklungen und Bedürfnissen anzupassen.

Vor diesem Hintergrund stellen sich bei der Konzeption eines geisteswissenschaftlichen Forschungsdatenzentrums, das sowohl die Langzeitarchivierung als auch die Bereitstellung der Forschungsdaten für die Nachnutzung sicherstellen soll, verschiedene Fragen:

- Was sind Forschungsdaten in den Geisteswissenschaften und welche Forschungsdatentypen sollen vom Angebot des Datenzentrums berücksichtigt werden? Wie können geeignete Objektmodelle für die Bereitstellung und Archivierung dieser Forschungsdaten aussehen? Wie kann mit Forschungsdaten umgegangen werden, die nicht dokumentenbasiert sind, sondern beispielsweise aus Datenbanken bestehen?

- Wie kann die Zusammenarbeit zwischen Wissenschaft und Forschungsdatenzentrum erfolgreich sein? Welche Angebote hinsichtlich Beratung und Schulung sind besonders geeignet, um der Bedeutung des Forschungsdatenmanagements gerecht zu werden? Welche Implikationen hat das für mögliche Organisationsformen (-einheiten) eines Forschungsdatenzentrums?
- Die langfristige Nachhaltigkeit und Nachnutzbarkeit von Forschungsdaten ist nicht nur ein technisches, sondern vor allem ein organisatorisches Thema. Bestimmte geisteswissenschaftliche Forschungsdaten (zum Beispiel Editionen, Korpora, Wörterbücher) behalten über einen längeren Zeitraum ihre Forschungsrelevanz. Wie lässt sich diese Anforderung in organisatorischer und infrastruktureller Hinsicht umsetzen?
- Welche bestehenden und zukünftigen Standards für ein Forschungsdatenzentrum sind zu beachten, um Interoperabilität und Kooperation zwischen Forschungsdatenzentren zu fördern? Wie kann dies in Einklang mit der Anforderung der Skalierbarkeit gebracht werden?
- Ein Forschungsdatenzentrum muss über einen längeren Zeitraum lernen und sein Angebot anpassen. Gleichzeitig ist aber die Stabilität der konkreten (technischen) Angebote wichtig: für die technische Infrastruktur zum stabilen Aufbau der technischen Dienste; für Nutzer des Forschungsdatenzentrums um bereits zu Projektbeginn die Angebote in ihr Datenmanagement einplanen zu können und auch zu Projektende noch darauf vertrauen zu können. Stabilität und Innovation sind dabei zwar keine Gegensätze, müssen aber gegeneinander abgewogen werden. Wie leistet ein Forschungsdatenzentrum diesen Ausgleich zwischen Erneuerung und Stabilität des Angebotes?

Ein Wizard für die Erschließung strukturierter Textdaten

Fritz Kliche¹, Nicolas Schmidt², Ulrich Heid¹

¹Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim

²Institut für Betriebswirtschaft und Wirtschaftsinformatik, Universität Hildesheim
{kliche,schmi032,heid}@uni-hildesheim.de

Wir stellen einen *Wizard* vor, mit dem strukturierte Textdaten in einer Browser-Anwendung erschlossen werden können, um die textlichen Inhalte und Metadaten für textwissenschaftliche Analysen nutzen zu können. Der *Wizard* ist ein interaktives Werkzeug, das es dem Benutzer erlaubt, von Beispielfällen auf größere Mengen von Daten zu generalisieren, ohne dass er dazu zu programmieren braucht.

Die Voraussetzung sind Textdaten, deren Textstruktur nicht für jeden Text unterschiedlich ist, sondern wo sich ähnliche Textstrukturen über größere Mengen von Einzeltexten hinweg beobachten lassen. Beispiele sind Sammlungen von Zeitungsartikeln, Sammlungen von Blogs und user-generated content oder Protokolle von Parlamentsdebatten. Solche Texte sind einerseits nicht standardisiert, andererseits doch relativ homogen repräsentiert, mindestens innerhalb jeder Kollektion, jedes Zeitungs- oder Blog-Archivs. Die extrahierten Inhalte werden als *Textobjekte* in einer Datenbank abgelegt und mit Labels versehen, die den Zugriff ermöglichen. Über diesen Zugriff können die Textobjekte in eine neue Datenstruktur überführt werden. Der *Wizard* entsteht innerhalb des DH-Projekts *e-Identity* (Blessing et al., 2013), in dem ein umfangreiches Sample von Zeitungsartikeln (>800.000 Artikel) aus 5 digitalen Medienportalen erschlossen und ein Korpus erstellt wurde, in dem die textlichen Inhalte der Artikel und die begleitenden Metadaten kategorisiert vorliegen.

Der *Wizard* führt den Anwender durch die Funktionen, die über eine Browser-GUI gesteuert werden. Zunächst können Textdaten in unterschiedlichen Formaten (RTF, DOCX, ODT, TXT, HTML) und Zeichencodierungen (UTF-8, ISO-8859-1) importiert werden. Die anschließende Erschließung erfolgt in zwei Schritten: (1) Die Textdaten werden zunächst in strukturelle Einheiten (z. B. in *e-Identity*: Zeitungsartikel) segmentiert; (2) in diesen Einheiten werden anschließend textliche Inhalte und Metadaten erkannt und klassifiziert, indem ihre Anfangs- und Endpunkte im fortlaufenden Textmaterial identifiziert werden. Dafür werden in einem Vorschau-Fenster Ausschnitte der importierten Textdaten angezeigt. Der Anwender erstellt anhand solcher Beispiele *Extraktionsregeln*, d. h. Muster, nach denen Textobjekte identifiziert werden. Mit der Erstellung mehrerer Extraktionsregeln entsteht ein Regelset als eine Schablone, mit der in den importierten Daten Inhalte erkannt und extrahiert werden. Für die Extraktionsregeln wurden Elemente

einer Regelsprache implementiert, über die der *Wizard* computerlinguistische Konzepte (reguläre Ausdrücke, Text Mining, computerlinguistische Prozessierung) textwissenschaftlichen Anwendern möglichst intuitiv zugänglich macht. Die im Folgenden dargestellten Konzepte wurden umgesetzt:

Integrierte computerlinguistische Werkzeuge

Der *Wizard* integriert computerlinguistische Verarbeitungsschritte zur Tokenisierung, Lemmatisierung, Wortartenerkennung und zur Erkennung von Eigennamen. Weiter werden Tokens verschiedenen „Tokentypes“ zugeordnet (Versalien, groß- oder kleingeschriebene Wörter, Zahlwörter usw.). Die entsprechende computerlinguistische Verarbeitung soll einerseits im Hintergrund stattfinden; andererseits soll sie von den Anwendern gesteuert werden können. Wir trennen dazu die Erstellung der Extraktionsregeln von ihrer Anwendung. Nach der Erstellung einer Schablone validiert der *Wizard* deren Regeln und prüft, welche computerlinguistischen Vorverarbeitungsschritte sie verlangen. Der Anwender muss also nicht entscheiden, welches computerlinguistische Werkzeug an welcher Stelle zum Einsatz kommen soll, sondern welches Prozessierungsergebnis angestrebt wird. Wo nötig, wird dazu ein computerlinguistischer Verarbeitungsschritt vom Werkzeug vorgeschlagen und eingeschoben.

Unterschiedliche Textobjekte

In den Daten können unterschiedliche Textobjekte definiert werden. Als Textobjekte sind Tokens, Segmente (d. h. einzelne Zeilen) und mehrzeilige Objekte möglich.

Merkmale zur Identifikation

Zur Erstellung der Extraktionsregeln können unterschiedliche textliche Merkmale berücksichtigt werden. Als Indikator eines Textobjekts können (1) ein Ankerwort oder (2) ein regulärer Ausdruck definiert werden; (3) die maximale und die minimale Länge eines Segments können festgelegt werden; (4) um das Segment im Kontext zu definieren, können Ankerwörter und reguläre Ausdrücke zum Vorgänger- oder Nachfolgersegment des zu bestimmenden Segments definiert werden. (5) Schließlich kann die Abfolge unterschiedlicher Typen von Tokens definiert werden, die über die Wortart, einen Abgleich mit Terminologielisten (z. B. eine Liste von Monatsnamen), „Tokentypes“ oder über eine feste Zeichenkette charakterisiert werden.

Funktionen der Extraktionsregeln

Die Extraktionsregeln dienen zunächst der Identifikation von Textobjekten; sie können auch als Blocker fungieren, die die Identifizierung von Objekten durch andere Regeln verhindern.

Der Aufbau einer eigenen Datenstruktur

Die erkannten Objekte werden in einer Datenbank mit ihrem Label abgelegt. Durch das Label kann auf die Daten zugegriffen werden. Damit sind die Daten für den Aufbau einer neuen Datenstruktur zugänglich. Die Daten können in ein XML-Format konvertiert werden. Wir planen die Möglichkeit zur Konvertierung in gängige Formate: TEI, CMDI (Broeder et al., 2012), etc.

Anwendung für Textwissenschaftler in DH-Projekten

Das Poster richtet sich besonders an textwissenschaftliche Anwender. Screenshots stellen die Arbeitsschritte zur Erschließung strukturierter Textdaten dar. Aus computerlinguistischer Sicht zeigt das Poster ein Beispiel, wie linguistische Annotationen in ein Softwareprojekt für die Digital Humanities eingebunden werden. In einer Demonstration können die Benutzerschnittstellen des Systems vorgeführt werden.

Literatur

Blessing, André; Sonntag, Jonathan; Kliche, Fritz; Heid, Ulrich; Kuhn, Jonas; Stede, Manfred (2013). Towards a tool for interactive concept building for large scale analysis in the humanities. In: *Proceedings des 7. Workshops „Language Technology for Cultural Heritage, Social Sciences, and Humanities“*. Association for Computational Linguistics, Sofia, Bulgarien.

Broeder, Daan; Windhouwer, Menzo; van Uytvanck, Dieter; Goosen, Twan; Trippel, Thorsten (2012). CMDI: a component metadata infrastructure. In: *Proceedings des Workshops „Describing Language Resources with Metadata“*. LREC 2012, Istanbul, Türkei.

Database of historical places, persons and lemmas

Natalia Korchagina ^{1, 2}

¹ *Schweizerische Rechtsquellenstiftung / Zürich, Switzerland*

² *Institut für Computerlinguistik / Universität Zürich, Switzerland*

Proposed session - Poster session

Keywords - digital humanities, database, RDF triple store, multilinguality, NLP for historical texts

The importance of the representation of humanities material as structured, interconnected objects has grown with the recent emergence of the ideas of Linked Data and Semantic Web. Efficient cataloging and storing of humanities data can facilitate the research and knowledge exchange within the field. In this respect, the use of modern technologies of data storage, i.e. databases, is a crucial point for a digital humanities project.

The Swiss Law Sources Foundation has been handling the critical edition and publishing of Swiss historical legal manuscripts for over a hundred years. By today, over 100 volumes of texts have been published, about 30 of them are available as digital editions. This collection contains texts in German, French, Italian, Romansh and Latin languages. The texts' creation time ranges from the 10th to the 18th centuries representing, for this reason, a rich source not only of historical, but also of linguistic information on language evolution. Each of these volumes contains a back-of-the-book index of persons, places and lemmas mentioned.

The creation of the database should facilitate the edition of the index of future volumes, as well as to be a starting point for users looking for information on a specific personality/place which could have been mentioned in several Foundation's volumes. The database will have the four CRUD (create, read, update, delete) basic functions of persistent storage, and will provide its users with an intuitive GUI. This database is intended for use in two directions: first, for the edition of indexes of the upcoming volumes where editors will update/create database entries via GUI instead of working with Excel files; and second, for large public browsing the database via GUI in read-only mode. To give some more details on a historical person/place the database entries will be linked (when possible) with the corresponding GND (Integrated authority file of the German National Library) and HLS (Historische Lexikon der Schweiz) entries.

The technology to be used is RDF triple store. This is a NoSQL (non-relational) mechanism for storing and retrieval of data. In a triple store each data entity is composed of subject-predicated-object (triple), like "John knows Mary". An RDF triple store can be viewed as a graph, where an object of one entity is a subject of another, and so on. Graph data representation is particularly pertinent for Digital Humanities where the data is highly interconnected. On practice this kind of data representation guarantees fast path-walking for complex queries enabling knowledge discovery. Furthermore, an RDF triple store has a simple and uniform standard data model, and is governed by a powerful standard query language SPARQL. An RDF triple store also provides a standardized interchange format (e.g. N-triples) for import/export which is important for data transfer/exchange. Thus, RDF triple store is a mature, stable technology convenient for persistent data storage. Moreover, RDF is a standard model for data interchange over the emerging Semantic Web and Linked Open Data cloud. As a future goal, we aim to integrate our RDF data into other international projects (e.g. DBpedia, Europeana) for a higher visibility. Another future direction would be participation in such "meta"-projects as, for example, Bibliographie-Portal (<http://www.biographie-portal.eu>).

Mit den Informationswissenschaften von Daten zu Erkenntnissen

Sandra Balck, Prof. Dr. Stephan Büttner, Denise Ducks, Ann-Sophie Lehfeld, Eva Schneider, Evelyn Vietze

Fachhochschule Potsdam, Fachbereich Informationswissenschaften

1. Transformationsprozess Wissen- Information

Der Transformationsprozess von Wissen zu Informationen ist ein originär informationswissenschaftliches Problem. In der informationswissenschaftlichen Betrachtung ist Wissen der Ausgangspunkt von Daten und Informationen. Informationen gehen demnach nicht, wie in der klassischen DIKW-Pyramide (Data-Information-Knowledge-Wisdom) angenommen, aus Daten hervor sondern werden durch einen doppelten Transformationsprozess aus Wissen generiert. Anstelle eines hierarchischen Modells wird eine funktionale Unterscheidung zwischen formal-syntaktischen, semantischen und pragmatischen Ebenen von Information vertreten¹. *(Dieser Transformationsprozess wird im Poster durch eine Grafik visualisiert.)*

2. Beitrag der Informationswissenschaften

Die Informationswissenschaften (IW) verfügen über Methoden welche es ermöglichen vorhandenes Wissen aus Informationsbeständen zu extrahieren. Auch in Digital Humanities (DH)-Projekten werden neue Daten u.a. mit Hilfe der Methoden der Informationswissenschaften generiert und neue Erkenntnisse gewonnen. Dies betrifft alle die Informationswissenschaften tangierenden Disziplinen (Linguistik, Informatik u.a.). Vorhandenes Wissen ist auf Grund interdisziplinärer Zusammenarbeit nicht mehr klar voneinander zu trennen und sollte es auch nicht sein. Dies erfordert eine Optimierung des Transformationsprozesses. Die Methoden der Informationswissenschaften bieten dazu ein geeignetes Toolset.

IW ist, nicht nur, wie z.B. im Drei-Sphären-Modell von DARIAH² angenommen, ein geisteswissenschaftliches Einzelfach, sondern bietet eine gemeinsame, disziplinübergreifende theoretische Grundlage für DH. Bereits Roberto Busa wies im Companion to Digital Humanities³ explizit darauf hin, dass der größte der drei Stränge der DH als "documentaristic" or "documentary", zu bezeichnen sei.

Bei den Bestrebungen für ein Referenzcurriculum im Rahmen der DARIAH-Initiative wurde die informationswissenschaftliche Ausbildung bisher kaum oder gar nicht wahrgenommen.

¹ vgl. Kuhlen (2013)

² vgl. ARIAH Working Papers (2013) Schreibmann, et al (2004)

³ vgl. Schreibmann, et al (2004)

Ein Vergleich der zentralen Themen der DH mit denen der IW zeigen jedoch große Übereinstimmungen:

Suchverfahren
Text Mining und Sprachverarbeitung
(Forschungs-)Datenmanagement
Fachspezifische Datenbanken
Fachinformation
Geographische Informationssysteme
digitale Bildverarbeitung
User studies
Hermeneutik (third current)
Digitale Edition und
Langzeitarchivierung

3. Module Studiengang Informationswissenschaften der Fachhochschule Potsdam

In der informationswissenschaftlichen Ausbildung der Fachhochschule Potsdam spielen die Kernkompetenzen der DH eine wesentliche Rolle (siehe Tabelle).

Module	Credit Points
Erschließung	15-25
Datenbanken	5-25
Information Retrieval	5-20
Digitale Editionen	7
Dokument	10
Informationsvisualisierung	6
Modellierung / XML	10
Linguistik / Textmining	5
Datenmining / Semantic Retrieval	14
Wissenschaftsmethodik	5

Die Ausprägung der einzelnen Module unterscheidet sich hierbei innerhalb der drei angebotenen Studiengänge Archiv, Bibliothekswissenschaft sowie Information und Datenmanagement.⁴

4. DH als Anwendung informationswissenschaftlicher Methoden

Der drei-semestrige Masterstudiengang Informationswissenschaften bietet eine fachwissenschaftliche Weiterführung informationswissenschaftlicher Grundlagen mit zwei vertiefenden Profilierungsmöglichkeiten und baut auf einem informationswissenschaftlichen Bachelorstudium auf. Inhaltlich ist eine

⁴ Vgl. Studien- und Prüfungsordnung (2014)

Spezialisierung durch die Wahl einer von zwei Profillinien im zweiten Semester möglich.

Profil 1: Records Management und Digitale Archivierung

Profil2: Wissenstransfer und Projektkoordination

Die Profillinie "Wissenstransfer und Projektkoordination" vermittelt dabei Kompetenzen, die sich mit den Kerninhalten der DH überschneiden.

Im Track Wissenstransfer ist es demzufolge notwendig und folgerichtig DH als Anwendung informationswissenschaftlicher Methoden zu integrieren. Die Integration der Digital Humanities soll dabei nicht durch die Unterbringung neuer Inhalte, sondern die namentliche Verankerung (Implizites explizit machen) und somit Sichtbarmachung der bereits als IW-Methoden im Curriculum verankerten Kompetenzen erfolgen. IW fungiert dabei als Mittler zwischen D (Informatik) und H (Geisteswissenschaften).

Im Bestreben eine gemeinsame Lobby für den geisteswissenschaftlichen Transfer zu etablieren, wird keine Fusion sondern die Kooperation beider Disziplinen angestrebt. Diese könnte sich unter anderem in einer gemeinschaftlichen Ausbildung niederschlagen.

Literatur

DARIAH-DE Working Papers 2013-1

Sahle, P.: Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities

nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2013-1-5

Kuhlen, R.: A1 Information – Informationswissenschaft

in: Kuhlen, R.; Semar, W.; Strauch, D. (Hrsg.): Grundlagen der praktischen Information und Dokumentation. 6. Ausgabe. Berlin 2013: Walter de Gruyter

Schreibmann, S.; Siemens, E.; Unsworth, J. (Ed): A Companion to Digital Humanities Oxford, Blackwell (2004)

Studien- und Prüfungsordnung für die Bachelorstudiengänge

Archiv, Bibliothekswissenschaft, Information- und Datenmanagement des Fachbereichs Informationswissenschaften der Fachhochschule Potsdam - Besondere Bestimmungen (2014 intern)

Poster-Bewerbung. DHd 2015, Graz – „Von Daten zu Erkenntnissen“**Themenbereich:** (a) Mehrwert digitaler Methoden und Technologien für Erkenntnisprozesse**Arbeitstitel:** Auch ich in Rom! Die literarische Inszenierung sozialer Netzwerke und Wissenstransfers in deutschsprachigen Reiseberichten (1816-1833)**Fachbereich, Universität:** Neuere Deutsche Literaturwissenschaft, Humboldt-Universität zu Berlin**Betreuer:** Prof. Dr. Steffen Martus, Prof. Dr. Anne Baillot

Zusammenfassung: Das Rom des 19. Jahrhunderts ist auch die Stadt der Bildungsreisenden. Villen, Cafés und Galerien fungieren als Treffpunkte; Erlebnisse und Fachwissen aus Kunstgeschichte, Mineralogie, Philosophie etc. werden ausgetauscht. Das nachfolgend beschriebene Forschungsprojekt analysiert die literarische Inszenierung sozialer Netzwerke und Wissenstransfers in Rom, dargestellt in deutschsprachigen Reiseberichten (1816-1833). Zur Anwendung kommen computergestützte Methoden: Die Berichte werden in XML semantisch ausgezeichnet, die gewonnenen Daten sollen daraufhin in eine Datenbank eingespeist und visualisiert werden. Im zweiten Teil der Untersuchung werden diese unter literatursoziologischen und wissenspoetologischen Gesichtspunkten interpretiert.

Untersuchungsgegenstand und Zielsetzung: Im deutschsprachigen Raum kommt es seit den aufgeklärten Lesegesellschaften und der späteren Salon-Kultur zu zahlreichen Vereinsgründungen. Der Habitus des geselligen Gelehrten wird auch von Reisenden in Rom gepflegt: Intellektuellenzirkel, wie der Deutsche Künstlerverein oder die Zusammenkünfte im Antico Caffè Greco, entstehen. Fachwissenschaftliches Wissen wird gemeinsam erarbeitet und diskutiert – ‚Netzwerken‘ ist wesentlicher Bestandteil der Aufenthalte. Dieses Phänomen spiegelt sich in seiner Dynamik besonders in Reiseberichten wider: Goethe fährt nach Rom, beschreibt abendliche Malzirkel (Goethe [HA] 1974: 134-136) etc. Die dänisch-deutsche Schriftstellerin Friederike Brun wohnt mit dem Ehepaar von Humboldt zusammen und berichtet Begegnungen und philosophische Gespräche (Brun 1833: 171-179). Diese und andere literarisierte Begegnungen **sollen im dargelegten Projekt erstmalig umfassend analysiert und visualisiert werden.** Die Forschungsfrage lautet: **Welche sozialen Netzwerke werden in den Reiseberichten inszeniert und welche gemeinsam erörterten Diskurse prägen die Zusammenkünfte?**

Forschungskontext: Das Projekt ordnet sich in einen äußerst aktuellen Forschungskontext über Reisenetzwerke im 19. Jahrhundert ein: **Geschichte:** DHI Rom: *Künstler, Agenten und Sammler in Rom 1750-1850*; **Kunstgeschichte:** Karl S. Rehberg: SFB 804 *Transzendenz und Gemeinsinn* (Netzwerke deutscher und französischer Künstler in Rom); **Musikwissenschaften:** *Europäische Musiker in Venedig, Rom und Neapel (1650-1750)*.¹ Eine Analyse der Reiseberichte unter literaturwissenschaftliche Perspektive wurde bisher noch nicht vorgenommen. Diese Lücke soll nun geschlossen werden. Methodisch orientiert sich das Dissertationsvorhaben dabei insbesondere an dem oben genannten musikwissenschaftlichen Projekt. Damit wird der inhaltliche sowie methodische Anschluss an das Forschungsumfeld gewährleistet beziehungsweise wird dieses durch die literaturwissenschaftliche Perspektive angereichert. Die Einbindung der Forschungsergebnisse in eine Onlineplattform soll die

¹ Projekt-URLs siehe Literaturverzeichnis.

öffentliche Nutzung und Langzeitverfügbarkeit der Daten gewährleisten (Shillingsburg 2006: 12). Eine CC-BY-Lizenz wird bevorzugt, um Verbreitung zu ermöglichen.

Methodik und Vorgehen: Digitale Methoden kommen zur Anwendung, die einen **neuen Blick auf die Vielfalt der Reisebeschreibungen** eröffnen sollen. Die Strukturierung und Visualisierung können Interferenzen sichtbar machen, die durch eine klassische Textanalyse in diesem Umfang kaum sichtbar wären. Das zugrundeliegende **Korpus** wird aus ca. 30 Texten bestehen, eingeleitet durch Goethes *Italienische Reise* (1816) und beschlossen durch Friederike Bruns *Römisches Leben* (1833). Dieser Zeitraum ist ein Höhepunkt in der Italien-Reiseliteratur, es werden besonders viele Berichte verlegt.

Die meisten relevanten Berichte liegen bereits als PDF vor. Diese werden mit Hilfe der OCR-Erkennungssoftware ABBYY FineReader maschinenlesbar gemacht und in **XML** ausgezeichnet: Den Personen sind Attribute wie Beruf, regionale Herkunft, persönliches Verhältnis, Gesprächsthemen etc. zuzuordnen. Wo vorhanden, kommen **Normdaten-Identifizierungen** (GND-Referenzen) zum Einsatz. In einem nächsten Schritt werden die jeweiligen Personennetzwerke mit Gephi **visualisiert**,² die Datenmenge wird damit leichter interpretierbar. Zudem wird angestrebt, eine zeitliche Dimension mit Hilfe einer Timeline einzubauen, sodass eine dynamische Darstellung der Daten gewährleistet wird. Hierauf aufbauend sind die erhobenen **Daten auszuwerten**. Dazu werden Theorien aus dem Bereich der Literatursoziologie sowie der Wissenspoetologie herangezogen.³

Relevante Informationen auf dem Poster:

- Untersuchungsgegenstand, Forschungsfrage, exemplarisches Textbeispiel
- Methodik, Vorstellung verwendeter Tools
- Eigener Forschungsstand zum Zeitpunkt der Tagung (veranschaulicht in einem Projekt-Zeitstrahl)

Zitierte Literatur und Projekte:

- **Baßler**, Moritz (Hrsg.): New Historicism. Tübingen ²2001.
- **Brun**, Friederike: Römisches Leben. Band 1, Leipzig 1833.
- **Europäische Musiker in Venedig, Rom und Neapel (1650-1750)**: www.musici.eu [29.10.2014].
- **Goethe**, Johann W. von: Italienische Reise, in: Goethes Werke. Hrsg. von Erich Trunz. Hamburger Ausgabe. Band 11: Autobiographische Schriften III. Hamburg ⁸1974.
- **Hogrebe**, Wolfram: Societas Teutonica. Erlangen/ Jena 1996.
- **Klausnitzer**, Ralf: Literatur und Wissen. Berlin 2008.
- **Künstler, Agenten und Sammler in Rom 1750-1850**: <http://dhi-roma.it/projekte-aktuell+M5a1e59c05e9.html> [29.10.2014].
- **Rehberg**, Karl-Siegbert: SFB 804 Transzendenz und Gemeinsinn. www.sfb804.de [29.10.2014].
- **Shillingsburg**, Peter L.: From Gutenberg to Google. Cambridge 2006.
- **Vogl**, Joseph (Hrsg.): Poetologien des Wissens um 1800. München 1999.

² Siehe <http://gephi.github.io/> [29.10.2014].

³ Siehe bspw. Hogrebe 1996. Zum New Historicism siehe Baßler 2001. Zur Theorie einer „Poetologie des Wissens“ Klausnitzer 2008: 169-183 und Vogl 1999.

Digitalisierung eines NS-Bildarchivs – Konstruktion von NS-Lebenswelt

Posterpräsentation

Unser Projekt repräsentiert das Erste dieser Art im Feld der Theaterwissenschaft und unsere Intention ist es weitere Initiativen für unsere Disziplin anzuregen. Das sogenannte Bildarchiv ist Teil des Archivs und der historischen Theatersammlung des Instituts für Theater-, Film- und Medienwissenschaft der Universität Wien (TFMA). Im April 2011 wurde dieses verschollen geglaubte, umfangreiche historische Bildarchiv des ehemaligen „Zentralinstituts für Theaterwissenschaft“ wiederaufgefunden. Der Bestand umfasst vorwiegend Fotografien (Schauspielerporträts und Theaterfotografien zwischen 1880–1945), Stiche und Grafiken aus dem 19. Jahrhundert, insgesamt ca. 2000 Einzelstücke. Den Hauptteil bildet die visuelle Dokumentation von NS-Theatern, die in solcher Vollständigkeit keine österreichische oder deutsche Sammlungsinstitution aufzuweisen hat. Es handelt sich hierbei um Fotografien sämtlicher Produktionen an Wiener Bühnen im Zeitraum von 1938 bis zur Theatersperre im Juli 1944, von Prager Bühnen und Produktionen sogenannter Grenzlandtheater. Dieses Fotomaterial wurde dem Institut von verschiedenen Pressefotographen zur Verfügung gestellt. Weitere historische Fotos sind frühe und äußerst rare Schauspielerporträts, sehr häufig mit handschriftlichen Widmungen versehen. Der Aufbau dieses Bildarchivs war eine der ersten Maßnahmen des 1943 an der Universität Wien gegründeten „Zentralinstituts für Theaterwissenschaft“. In Kontext gesetzt zu den Zielvorgaben seitens des Reichserziehungsministeriums, nämlich nach dem Krieg als Reichsinstitut die Bedeutung von Theater und Film für das großdeutsche Reich zu definieren, hat dieser Fotobestand große Brisanz.

Dieses Bildarchiv bietet eine exzellente Möglichkeit zur Entwicklung einer Digital Humanities-Strategie für unser Fach und erleichtert fächer- und institutionenübergreifende Zusammenarbeit und Vernetzung. Unser Projekt zielt auf einen intensiven Austausch zwischen verwandten wissenschaftlichen Feldern ab. Die Interdisziplinarität ist eine wichtige Säule des Projekts, sowohl technisch als auch wissenschaftlich motiviert.

Dabei soll vorrangig auf bereits bestehende Tools und Standards zurückgegriffen werden, wobei an manchen Stellen auch Eigenentwicklungen notwendig sein werden. Diese werden idealerweise in bereits aktiv genutzte Strukturen einfließen. Zudem wird mit der Digitalisierung des Bildarchivs die strukturelle Grundlage geschaffen, um den gesamten Bestand des TFMA digital aufzubereiten. Es wird dafür ein für die Theaterwissenschaft prototypischer Workflow entwickelt werden.

Für die Präsentation und Bearbeitung der digitalisierten Objekte wird eine Webplattform erstellt, die die wissenschaftlichen Ergebnisse unseres Projekts zu Beginn in den Mittelpunkt stellt, dabei aber für zukünftige Forschungsarbeiten offen und jederzeit andockbar bleibt.

Die digitalisierten Objekte werden der Forschungsgemeinschaft und Öffentlichkeit in der freien Lizenz CC BY 4.0 zur Verfügung gestellt und in einem Open Access-Format angeboten. Durch die Wahl dieser Lizenzierung ist die Basis für eine breite (Nach-)Nutzung gelegt. Zudem wird auf Langzeitarchivierung Rücksicht genommen, indem die Digitalisate im Repositorium von PHAIDRA (Digital Asset Management System mit Langzeitarchivierungsfunktionen der Universität Wien, <http://phaidra.univie.ac.at>) eingebunden werden. Zudem werden mit Tools auf der Webplattform neuartige Zugriffe auf den Bestand möglich sein, der ein innovatives, zeitgemäßes wissenschaftliches Arbeiten unterstützt.

Mit den Bedeutungseinschreibungen zu Theater und Film lassen sich nicht alleine ästhetische Fragekomplexe aufwerfen. Noch dringlicher stellen sich dabei Fragen nach NS-Menschenbildern. Konstruktionsvorgänge dieser Menschenbilder werden über Theater- und Filmproduktion erkennbar. Die Digitalisierung dieses NS-Bildarchivs birgt für die internationale Forschung unterschiedlicher Disziplinen großes Innovationspotential, da sich über diese Materialien nicht allein Repräsentationsformen untersuchen lassen, sondern auch sichtbar wird, wie parallel zur Vernichtung der als „nichtarisch“ gekennzeichneten Personen ein neues NS-Menschenbild konstruiert wird.

Auf unserem Poster skizzieren wir die eben beschriebenen Abläufe exemplarisch anhand von ausgewählten Theaterfotografien aus diesem NS-Bildarchiv, um aufzuzeigen, wie eine digitalisierte Aufbereitung ideologische Sammlungsstrategien und Wissenschaftspolitik sichtbar macht. Die einzelnen Schritte werden sowohl auf technischer als auch inhaltlicher Ebene erläutert.

Projektteam

PD Mag. Dr. Birgit Peter
Mag. Klaus Illmayer
Mag. Johannes A. Löcker

Archiv und theaterhistorische Sammlung (TFMA)
tfm | Institut für Theater-, Film- und Medienwissenschaft
Universität Wien
Hofburg, Batthyanystiege
1010 Wien

Kontakt: birgit.peter@univie.ac.at



JENSITES – Schriften einer Totenbruderschaft digital

Das Poster **JENSITES – Schriften einer Totenbruderschaft digital** informiert über ein kleines Forschungsprojekt (www.oeaw.ac.at/iclitt/bruderschaftsdrucke), das derzeit – von der Stadt Wien gefördert – am Institut für Corpuslinguistik und Texttechnologie durchgeführt wird.

Zum Hintergrund: Bruderschaften waren christliche Vereinigungen, die im Zuge der Gegenreformation einen neuen Aufschwung erfahren und die religiöse Alltagskultur in ganz Europa stark geprägt haben. Während den barocken Bruderschaften als Forschungsgegenstand im internationalen Vergleich ein wachsendes Interesse entgegengebracht wird („Society for Confraternity Studies“), steht ihre Erforschung im deutschsprachigen Raum, von wenigen Einzelstudien abgesehen, erst am Anfang.

Das genannte Projekt hat daher den Anspruch, anhand einer in Wien gegründeten, sogenannten **Totenbruderschaft** zu zeigen, wie man sich diesem vernachlässigten, in höchstem Maße interdisziplinären Forschungsgegenstand mit digitalen Methoden nähern kann:



1) Sichtung und Digitalisierung der Quellen

Die Quellen der kaiserlich-königlichen Totenbruderschaft befinden sich in verschiedenen kirchlichen und öffentlichen Archiven und Bibliotheken. Die Bruderschaft hat einerseits Handschriftliches wie etwa Gründungsdokumente, Urkunden, Verträge und interne Aufzeichnungen hinterlassen und andererseits Gedrucktes wie etwa Statuten, Gebetsbücher für Mitglieder, Neujahrskalender, Predigten und Memento mori-Dichtung publiziert. Durch den Digitalisierungsprozess wird dieses dislozierte Quellenmaterial virtuell zusammengeführt.

2) Aufbereitung und Annotation der Quellen

Um das reichhaltige Quellenmaterial zu erschließen, werden die gescannten Image-Digitalisate der Originale in maschinenlesbare Volltextversionen umgewandelt und in ein XML-Format (Version P5) überführt. In einem weiteren Schritt werden die Textdaten in einem semi-automatischen Verfahren mit linguistischen Informationen versehen (Tokenisierung, Wortklassenzuordnung und Lemmatisierung). Das Projekt profitiert hier von bereits abgeschlossenen Projekten, in denen am Institut für Corpuslinguistik und Texttechnologie daran gearbeitet worden ist, das automatische Tagging (Treetagger) durch die Verwendung bereits korrigierter Daten aus dieser Zeit zu verbessern und das standardisierte *Stuttgart Tübingen TagSet* für die Sprachstufe des Älteren Neuhochdeutsch zu adaptieren. Das

gewonnene Textmaterial, das sorgfältig kollationiert und korrigiert wird, erweitert einerseits die Datenbasis von historischen Texten aus dieser Zeit und ermöglicht andererseits die kontinuierliche Weiterentwicklung eines stabilen Methodeninventars für nicht-kanonische Varietäten.

3) Virtuelle Kontextualisierung der Quellen

Da die Totenbruderschaft eine kaiserliche Gründung war und ihr Vorstand und ihre Mitglieder zum Teil dem hohen Adel angehörten, sollen die in den Texten vorkommenden Personennennungen mit den vorhandenen RDF-Datensätzen der "Deutschen Biographie" <http://www.deutsche-biographie.de/> verlinkt werden. In Kooperation mit der Wiener Stadt- und Landesbibliothek wird derzeit an Referenzierungsmöglichkeiten von Orten, Bauwerken und Institutionen mit der kürzlich präsentierten historischen Wissensplattform "Wien Geschichte Wiki" <https://www.wien.gv.at/wiki/> gearbeitet. Mit dem Einsatz verschiedener Textanalysetools werden sich sprachliche und inhaltliche Bezüge innerhalb der Quellengruppe nachweisen und visualisieren lassen, was eine wesentliche Voraussetzung zur Interpretation und Funktionsbeschreibung dieser Texte sein wird.

Das Projektvorhaben – die Erforschung der Quellen der Totenbruderschaft – verbindet philologische Expertise mit moderner Informations- und Kommunikationstechnologie. Es hat daher zwar eine germanistisch-philologisch-kulturwissenschaftliche Ausrichtung, ist aber allein durch die gewählte Methodik der digitalen Quellenaufbereitung **interdisziplinär** angelegt: Die Quellen, die darin beispielhaft erschlossen werden, sind nicht nur ein wesentlicher Beitrag zu sozial-, kultur- und alltagsgeschichtlichen Aspekten der Stadt Wien, sondern auch Forschungsgegenstand der Sprachgeschichte und der Theologiewissenschaft und ermöglichen prosopographische Studien und historische Netzwerkforschungen.

Die Totenbruderschaft ist nur eine von vielen barocken Bruderschaften, doch eignet sich die günstige, bislang kaum erforschte Quellenlage in besonderer Weise dazu, beispielhaft aufbereitet und analysiert zu werden. **Es ist daher das Ziel des Projekts, die erhaltenen Quellen der Bruderschaft mit zeitgemäßen Methoden zu erschließen und in einem breiteren kulturwissenschaftlichen Kontext auszuwerten.** Auf Basis des digital aufbereiteten Materials sollen erstmals fundierte Aussagen über Geschichte, Mitglieder, Tätigkeit, Funktion und kulturelle Bedeutung dieser religiös begründeten Sozietät formuliert werden.

Bei der Quellenaufbereitung wird die Archivierung der Texte bereits mitbedacht – nach Abschluss des Projekts sollen die Daten Teil der Sammlung ABaC:us - Austrian Baroque Corpus werden.

TTLab Preprocessor – Eine generische Web-Anwendung für die Vorverarbeitung von Texten und deren Evaluation

Rüdiger Gleim und Alexander Mehler

Goethe-Universität Frankfurt

1 Einführung und Motivation

Dieser Beitrag stellt den *TTLab Preprocessor* (kurz: *TTLab PrePro*) als generische Web-Anwendung für die Vorverarbeitung von Texten in den *Digital Humanities* vor. Er erörtert die Architektur des *TTLab PrePro*, exemplifiziert das von ihm anvisierte Nutzungsszenario und fasst seinen aktuellen Entwicklungsstand zusammen.

Die linguistische Vorverarbeitung von Texten ist ein integraler Bestandteil jeder automatischen Textanalyse. Dies beinhaltet unter anderem die Erkennung der dem jeweiligen Text zugrundeliegenden Sprache(n), die Erkennung seiner logischen Dokumentstruktur, die Tokenisierung und Lemmatisierung seiner lexikalischen Konstituenten und die Annotation ihrer Wortarten (*PoS-Tagging*). Es existiert eine Reihe von Software-Systemen und -Komponenten, welche die Vorverarbeitung für verschiedene Sprachen umsetzen. In der Literatur werden dabei etwa für das PoS-Tagging Erkennungsraten von über 95% dokumentiert.¹ Für viele Fragestellungen, wie z.B. die Textklassifikation, fällt eine entsprechende Fehlerquote von ca. 5% kaum ins Gewicht. Im Bereich der *Digital Humanities*, bei der es etwa um die qualitative Analyse einzelner Wortbedeutungen geht, sind jedoch bereits Fehlerquoten von 1% oftmals inakzeptabel.² Gerade in diesem Bereich ist die automatische Vorverarbeitung zumeist der Ausgangspunkt für die nachfolgende unabdingbare manuelle Korrektur der Annotationen.

So stellt sich die Frage etwa zu Beginn eines Forschungsprojekts, wie hoch die erwartete Fehlerquote für Texte der untersuchten Sprache beim Einsatz eines bestimmten Präprozessierers ist. Zur Beantwortung dieser Frage kann eine Sammlung von Texten manuell vorverarbeitet und als so genannter *Gold-Standard* zur Bewertung der automatischen Vorverarbeitung herangezogen werden. Vergleicht man die Annotationsergebnisse verschiedener Systeme mit einem solchen Goldstandard, so können Kennzahlen zur Ermittlung der erwarteten Fehlerrate gewonnen werden, um schließlich den Aufwand für entsprechende manuelle Korrekturen zu schätzen. Da die Parametrisierung sowie die Ein- und Ausgabeformate verschiedener Systeme zur Vorverarbeitung variieren, ist die Durchführung einer solchen Evaluation aufwendig und ihrerseits fehleranfällig. Die Funktion, verschiedene Systeme über eine generische Schnittstelle nicht nur verwendbar, sondern auch evaluierbar zu machen, bildet folglich den funktionalen Kern des *TTLab PrePro*.

¹Diese Rate schwankt erwartungsgemäß je nach Sprache und Genre der untersuchten Texte [Giesbrecht and Evert, 2009].

²Anne Bohnenkamp-Renken (2013); *persönliche Kommunikation*.

2 TTLab Preprocessor Web-Anwendung

Der *TTLab PrePro* ermöglicht die Vorverarbeitung von Texten, die automatische Evaluation auf der Basis von Goldstandards und die einzelfallbezogene Fehleranalyse. Die Eingabe in das System kann direkt über den Browser in Form einer Texteingabe, die Angabe einer Webressource oder den Upload von Dateien erfolgen. Die Upload-Funktion ermöglicht nicht nur das Hochladen mehrerer Dateien auf einmal, sondern auch die Verwendung von komprimierten Archiven. An Dateiformaten werden unter anderem HTML, PDF, RTF und DOC unterstützt. In der Voreinstellung wird die Sprache der Inputtexte automatisch erkannt und der für die jeweilige Zielsprache voreingestellte Präprozessor verwendet. Es ist auch möglich, diese Parameter explizit zu setzen. Die Ausgabe erfolgt mittels *TEI P5* [TEI, 2014]. Die Ergebnisse können direkt im Browser in verschiedenen Sichten betrachtet und frei heruntergeladen werden.

Werden TEI-P5-Dokumente als Eingabe verwendet, so werden diese vom System – wie bei jedem anderen Eingabeformat – auf den unstrukturierten Text heruntergebrochen. Anschließend werden sie durch den Präprozessor vorverarbeitet und in TEI P5 repräsentiert. Bilden annotierte TEI-P5-Dokumente den Input, so können diese als Goldstandard interpretiert werden. Das System evaluiert in diesem Falle den jeweils ausgewählten Präprozessor auf der Basis dieses Goldstandards. Da die Tokenisierung zwischen den zu vergleichenden Dokumenten variieren kann, wird zunächst mittels dynamischer Programmierung ein Alignment der Token durchgeführt. Anschließend wird das Ergebnis der Lemmatisierung sowie des Taggings mit dem Goldstandard verglichen. Auf diese Weise können die aus dem *Machine Learning* bekannten Maße *Precision*, *Recall* und *F-Score* berechnet werden. Die Ergebnisse werden direkt im Browser angezeigt – sowohl für die einzelnen Dokumente, als auch für das Eingabekorpus insgesamt. Analog wird eine Rangverteilung der häufigsten Tagging- und Lemmatisierungsfehler (nach abnehmender Häufigkeit) visualisiert. Schließlich können die Tagging- und Lemmatisierungsfehler in einer tabellarischen Ansicht im jeweiligen Satzkontext untersucht werden. Abbildung 1 exemplifiziert eine solche Ansicht von Evaluationsergebnissen. Die obere Tabelle beinhaltet eine Liste aller evaluierten Dokumente mit den Gesamtergebnissen. Für ein ausgewähltes Dokument können, wie in diesem Beispiel gezeigt, Belegstellen von Tagging-Fehlern im Satzkontext aufgezeigt werden. Dies erlaubt das gezielte Nachverfolgen und Beheben von Fehlern.

Der *TTLab PrePro* ist als Java- und JavaScript-basierte Client-Server-Architektur implementiert. Die Benutzeroberfläche ist mithilfe des JavaScript-Frameworks ExtJS realisiert. Das in *Apache Tomcat* laufende *Java Servlet* bearbeitet die Nutzeranfragen, ruft externe Systeme zur Vorverarbeitung auf, führt ggf. Evaluationen durch und bereitet die Ergebnisse für die Darstellung im Browser auf. In der aktuellen Version sind zwei Systeme des *TTLab Preprocessor* [Mehler et al., 2015, Waltinger, 2010] integriert sowie das System namens *Stanford CoreNLP* [Manning et al., 2014].

3 Zusammenfassung und Ausblick

Der vorliegende Beitrag stellt den *TTLab PrePro* als System zur Vorverarbeitung von Texten und darauf basierenden Evaluationen vor. Das mit der geplanten Publikation veröffentlichte System ist frei ver-

The screenshot shows the TITLab PrePro web interface. At the top, there is a 'Preprocessing' section with an 'Options' dropdown and a text input field 'Enter text to preprocess here'. Below this are navigation tabs: Home, Preprocessing, Preprocessors Overview, Documentation, Technologies, Publications, Support, and Impressum. The main content area is divided into two panes. The left pane, titled 'Preprocessed Documents', shows a table with columns 'Docum...', 'Language', 'Tokens', and 'Distinct...'. The right pane, titled 'Evaluation Results', contains a table of 'Evaluated Documents' with columns for 'Document', 'Tokens', 'microAvg Precision', 'microAvg Recall', and 'microAvg FScore'. Below this is a 'PoS Error Frequency Chart' and a 'Lemma Error Frequency Chart', followed by a 'PoS Error Table' and a 'Lemma Error Table'. The 'PoS Error Table' is the most detailed, showing columns for 'Evaluation', 'Reference', 'Frequency', 'Document', 'Left Context', 'Token', and 'Right Context'. The bottom of the interface shows a 'Page 2 of 2' indicator and 'Displaying 26 - 47 of 47' items.

Document	Tokens	microAvg Precision	microAvg Recall	microAvg FScore
House of Usher (Poe).xml	15066	0.9934513027727463	0.9934513027727463	0.9934513027727463
Corpus	15066	0.9934513027727463	0.9934513027727463	0.9934513027727463

Evaluation	Reference	Frequency ↓	Document	Left Context	Token	Right Context
NN	RB	5	House of Usher (Poe).xml	and a tremulous q...	habitually	characterized his utt...
NN	JJ	5	House of Usher (Poe).xml	but, feeling the rai...	gauntleted	hand;
NN	RB	5	House of Usher (Poe).xml	and now pulling th...	hollow	-sounding
JJ	VBD	4	House of Usher (Poe).xml	for the feeling was	unrelieved	by any of that , beca...
JJ	VBD	4	House of Usher (Poe).xml	for the feeling was...	unnerved	me in the contemplat...
JJ	VBD	4	House of Usher (Poe).xml	his eyes were	tortured	by even a faint light;
JJ	VBD	4	House of Usher (Poe).xml	He admitted, howe...	unmingled	with dread;long-conti...
VBG	JJ	4	House of Usher (Poe).xml	One of the phanta...	exceeding	depth below the surfa...

Abbildung 1: Ansicht von Evaluationsergebnissen, welche Tagging-Fehler mit Belegstellen im Satzkontext aufzeigt.

wendbar (*open access*). Die Weiterentwicklung zielt auf die Nutzbarmachung des UIMA-Frameworks³. Zum einen, um den Pool der verfügbaren Systeme zur Vorverarbeitung zu vergrößern, zum anderen, um umfangreiche Parameterstudien über die einzelnen Komponenten durchführen zu können. Ferner soll eine Normalisierung von PoS-Tagsets für die Evaluation entwickelt werden. Der *TITLab PrePro* zielt vor allem darauf, von Geisteswissenschaftlerinnen und -wissenschaftlern auch ohne Informatik-Vorkenntnisse genutzt werden zu können. Unterstützt werden derzeit die Sprachen Latein [Mehler et al., 2015], Englisch und Deutsch.

Der *TITLab PrePro* kann unter der URL <http://prepro.hucompute.org> getestet werden. Ein TEI P5-Dokument zum Testen der Evaluation steht unter der Adresse <http://prepro.hucompute.org/examples/poe.tei> bereit.

Danksagung

Diese Arbeit ist im Rahmen des BMBF-Projekts *Computational Historical Semantics* (www.comphistsem.org) entstanden, für dessen Unterstützung wir uns herzlich bedanken.

Literatur

TEI P5: Guidelines for electronic text encoding and interchange, 2014. URL [\url{http://www.tei-c.org/Guidelines/P5/}](http://www.tei-c.org/Guidelines/P5/).

³<https://uima.apache.org/>

- Eugenie Giesbrecht and Stefan Evert. An evaluation of part-of-speech taggers for the web as corpus. In *Proceedings of DGfS-CL Postersession 2009*, 2009.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014. URL <http://www.aclweb.org/anthology/P/P14/P14-5010>.
- Alexander Mehler, Tim vor der Brück, Rüdiger Gleim, and Tim Geelhaar. *Towards a Network Model of the Coreness of Texts: An Experiment in Classifying Latin Texts using the TTLab Latin Tagger*. Theory and Applications of Natural Language Processing. Springer, Berlin/New York, 2015.
- Ulli Waltinger. *On Social Semantics in Information Retrieval*. Phd thesis, Bielfeld University, Germany, 2010.

Projekt Altägyptische Wörterbücher im Verbund: Digital unterstützte Analyse der Entwicklung ägyptischer Wörterbücher

In der Frühzeit der wissenschaftlichen Untersuchung des Ägyptischen im 19. und frühen 20. Jahrhundert ist eine Vielzahl von Wörterbüchern, Wortlisten und Glossaren entstanden. Da die Ägyptische Sprache erst allmählich entschlüsselt und verstanden wurde, entwickelte sich auch die Sicht auf die Erfassung der ägyptischen Wörter weiter. Alle aktuell bekannten Wörter des Ägyptischen sind im Thesaurus Linguae Aegyptiae (<http://aaew.bbaw.de/tla>, auch Berliner Wortliste, kurz BWL) erfasst. Um nachzuvollziehen, wie sich das Verständnis des Ägyptischen über die Zeit verändert hat, schafft das vorliegende Projekt eine Infrastruktur dafür, das Vorkommen von Wörtern in ägyptischen Wörterbüchern und anderen lexikographisch relevanten Publikationen mit den Einträgen der BWL zu verknüpfen. Mit dem entwickelten Werkzeug sind automatisierte Auswertungen der Entwicklung des Verständnisses der ägyptischen Lexik – und damit der Lexikographie – möglich, die ohne IT-Unterstützung nicht denkbar wären.

Das Projekt „Altägyptische Wörterbücher im Verbund“ orientiert sich an Wörterbuchportalen wie "Wörterbuchnetz" (<http://woerterbuchnetz.de/>), "OWID" (<http://www.owid.de/>), "Etymologiebank" (<http://www.etymologiebank.nl>) und Infolux (<http://infolux.uni.lu/worterbucher/>). Spezifisch für das Ägyptische ist, dass die meisten Wörterbücher handgeschrieben sind und neben deutschen, französischen, englischen oder italienischen Textteilen auch Hieroglyphen und Partien in Demotisch, Koptisch, Griechisch, Lateinisch, Hebräisch und Arabisch enthalten (können), ganz abgesehen von Transliterationen mit Zeichen, die nicht alle im Unicode-Format existieren. Unter diesen Umständen ist eine Texterfassung per OCR und eine anschließende Auszeichnung in TEI/XML nicht möglich, so dass Verknüpfungen von Bilddaten mit einer Lemmaliste vorgenommen und Metadaten an diese Verknüpfungen angehängt werden müssen.

Die unterschiedlich strukturierten (teils nach hieroglyphischer Orthographie, teils nach Transliteration, teils nach konventionalisierter hieroglyphischer Zeichenliste) und in der Methodik der Worttransliteration individuell verfahrenen Wörterbücher machen es unmöglich und unnötig, für jedes Wörterbuch eine eigene Lemma-Liste zu erstellen und diese mit einer Hyper-Lemma-Liste zu verknüpfen. Stattdessen wird die BWL als standardisierte Lemma-Liste eingeschaltet, was zudem eine zukünftige Verknüpfung mit der ägyptischen Textdatenbank ermöglicht. So werden die alten Wörterbücher langfristig gesehen mit einem System, das ägyptische Volltexte erfasst, verschränkt.

Zur Erfassung und Aufbereitung der Informationen über die Entwicklung ägyptischer Wörterbücher wurde ein webbasiertes Werkzeug erstellt. Dieses bietet eine für mehrere Forscher mit differenzierenden Fragestellungen gleichzeitig zugängliche Oberfläche, um die Verknüpfung von Wörterbuch- und anderen Publikationseinträgen mit Wörtern der BWL zu ermöglichen. Die BWL wurde in einem Vorverarbeitungsschritt extrahiert, so dass die Wörter mit Metadaten als Graphiken vorlagen. Die einzelnen Seiten der oft handgeschriebenen Publikation lagen ebenfalls als Graphiken vor. Metadaten zu den Graphiken und zur Verknüpfung wurden in einer MySQL-Datenbank gespeichert, auf die anschließend mit Hibernate zugegriffen wurde. Hierbei wurde bspw. die Möglichkeit geschaffen, zu jeder einzelnen Verknüpfung zu annotieren, wie zutreffend die damalige Transliteration und Übersetzung aus heutiger Sicht waren.

Um die Verknüpfung herzustellen, musste die Möglichkeit geschaffen werden, in einer Eingabeoberfläche zu erfassen, welcher Bereich einer Seite mit welchem Eintrag der BWL zusammenhängt. JSF und PrimeFaces enthalten bereits verschiedene Funktionalitäten, um diese Verknüpfung herzustellen, bspw. Datentabellen mit Suchfunktion für die Wörter der BWL und ein Verfahren, um Bildbereiche in den Publikationsseiten zu extrahieren. Aus diesem Grund wurden diese zur Umsetzung der Weboberfläche eingesetzt und so ein Werkzeug geschaffen, mit dem sich Wörter der BWL mit Publikationsseitenbereichen verknüpfen lassen.

Weiterhin wurde eine Möglichkeit geschaffen, die Publikationen nach festgelegten Kriterien zu analysieren. Hierbei wurden Anfrageoptionen nach Metadaten der einzelnen Verknüpfungen und statistischen Daten der Metadaten mit MySQL implementiert. Diese lassen sich bei Bedarf problemlos erweitern. Auf diese Weise lassen sich die großen Datenmengen der Wörterbücher schnell und effizient auswerten.

Mit dem beschriebenen Werkzeug wurden seitdem verschiedene ägyptische Wörterbücher des 19. Jahrhunderts erfasst und analysiert. Hierbei wurden in insgesamt zehn Publikationen 30371 Verknüpfungen erstellt. Die sukzessive erweiterbare Datenbank und die installierten Funktionen werden den Nutzer in die Lage versetzen, die einzelnen Etappen der Erforschung der ägyptischen Lexik und die jeweils zu Grunde gelegte Konzeption und Methodik besser nachvollziehen zu können. Im idealen Falle soll es möglich sein, die Forschungsgeschichte eines Lexems von den Anfängen der Ägyptologie bis zum Erscheinen des Wörterbuchs der Ägyptischen Sprache (1926-31), das die Basis der BWL darstellt, und darüber hinaus skizzieren zu können. Je nach Suchkriterium können dabei exemplarische bzw. statistische Daten ausgewertet werden, wie z.B. Qualität und Quantität der verwendeten Primär- und Sekundärquellen, Häufigkeit von lexikographischen Fehlern und Missdeutungen, die entweder mangels besserer Kenntnis und auch gelegentlich aufgrund fehlender Sorgfalt vorgekommen sind.

Insgesamt ermöglicht die so geschaffene IT-Infrastruktur eine Analyse der frühen ägyptischen Lexikographie, die sonst nicht bzw. nur durch sehr zeitaufwändige Recherchen möglich wäre. Die Präsentation wird sowohl die zugrundeliegenden Fragestellungen der Ägyptologie und bereits gefundene Lösungen als auch die technische Infrastruktur, die zur Unterstützung der Beantwortung der Fragestellungen geschaffen wurde, darstellen.

Ingo Börner, Angelika Hechtl

Quantitative Aufführungsanalysen zu Stücken Johann Nestroys

Die Posterpräsentation stellt einen Ansatz einer computergestützten quantitativen Aufführungsanalyse vor.

Basierend auf dem von Solomon Marcus (1970) vorgelegten mathematischen Dramenmodell, das durch Manfred Pfisters (2001) seine theoretische Fundierung sowie von Hartmut Ilseman (1890) eine praktische Umsetzung in der Analyse der Dramen von Shakespeare erfahren hat, werden im projektierten Vorhaben die Möglichkeiten zur Anwendung quantitativer Verfahren zur Analyse von Inszenierungen ausgelotet. In der bisherigen Anwendung quantitativer Verfahren auf Dramen stand der Dramentext alleine im Zentrum des Erkenntnisinteresses. Die konkrete Umsetzung als Inszenierung ist im Unterschied zum Film bisher nicht unter Rückgriff auf quantitative Methoden untersucht worden.

Die quantitative Aufführungsanalyse ermöglicht es, unterschiedliche Inszenierungen der Stücke Johann Nepomuk Nestroys anhand des Merkmals Bühnenpräsenz untereinander und mit dem Dramentexten zu vergleichen. Die entwickelte Methode wird exemplarisch anhand einiger ausgewählter Dramentexte („Der Talisman“, „Der böse Geist Lumpazivagabundus“ und „Der Zerrissene“) erprobt.

Untersuchungsgegenstand bilden sowohl aktuelle Aufführungen an Wiener Bühnen, als auch Aufzeichnungen von älteren Aufführungen (wie etwa Salzburger Festspielinszenierungen). Mit dem Merkmal „Bühnenpräsenz“ wurde in Anlehnung an das von Erika Fischer-Lichte (2007) beschriebene System theatralischer Zeichen als jenes Charakteristikum identifiziert, welches einen Vergleich von Inszenierungen untereinander und mit dem Dramentext ermöglicht. Es wird ermittelt, welche SchauspielerInnen zu welchem Zeitpunkt auf der Bühne anwesend sind. Die erhobenen Daten lassen sich mit dem Dramentext in Beziehung setzen.

Das erhobene Datenmaterial wird mittels R aufgearbeitet und visualisiert. So sollen im Text vorhandene Strukturen und ihre konkrete Realisierung in den unterschiedlichen Aufführungen nachvollziehbar gemacht werden.

Fischer-Lichte, E. (2007): *Semiotik des Theaters. Bd. 1. Das System der theatralischen Zeichen*. Tübingen.

Ilseman, H. (1898): *Shakespeare Disassembled. Eine quantitative Analyse der Dramen Shakespeares*. Frankfurt a. M.

Marcus, S. (1970): „Ein mathematisch-linguistisches Dramenmodell“. In: *Zeitschrift für Literaturwissenschaft und Linguistik*, S. 139–152.

Pfister, M. (2001): *Das Drama. Theorie und Analyse*. München.

Dr. Jakub Šimek

Universität Heidelberg
Germanistisches Seminar
Hauptstraße 207-209
D-69117 Heidelberg
+49-(0)6221-543217

jakub.simek@gs.uni-heidelberg.de

Christoph Forster

datalino. Forster, Fabian, Krumnow PartG
Martin-Luther-Straße 120
D-10825 Berlin
+49-(0)30-78893232

forster@datalino.de



**UNIVERSITÄT
HEIDELBERG**
ZUKUNFT
SEIT 1386

DHd-Tagung 2015

Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als
Mittler zwischen Information und Interpretation

Kodierung, Analyse und Visualisierung mittelalterlicher Kodexstrukturen im editorischen Kontext

Posterpräsentation – Abstract

Für die Beschreibung der Lagen- und Blattstruktur mittelalterlicher Kodizes wird in gängigen Handschriftenbeschreibungen meist die Chroust'sche Lagenformel verwendet. Diese zeigt durch römische Buchstaben die Anzahl der Doppelblätter in einer Lage an, während hochgestellte Ziffern auf die Blattzählung verweisen. Bei Wiederholung gleichartiger Lagen in einem Kodex wird die Art dieser Lagen zusammen mit deren Anzahl nur einmal angegeben. Fehlende Blätter werden nur durch Minuszeichen, eingefügte durch Pluszeichen und deren Anzahl bei einer Lage angedeutet, ohne dass normalerweise eine genaue Zuordnung zum konkreten Doppelblatt möglich wäre. Die TEI sieht für den Inhalt des Elements `<collation>` keine genauere Spezifizierung vor.

Eine Schwäche derartiger Beschreibungen ist, dass sie einerseits nicht ohne Weiteres maschinenlesbar und eindeutig genug sind, um mit digitalen Faksimiles verknüpft zu werden, und dass sie andererseits von den meisten Benutzern gedruckter und digitaler Ausgaben kaum wahrgenommen werden. Dabei ist die physische Lagen- und Blattstruktur gerade bei individuell hergestellten Buchartefakten wie mittelalterlichen Kodizes häufig wesentlich für das Verständnis der Textgestaltung und -überlieferung sowie der Seitenarrangements. So sind etwa Textlücken in einer Abschrift potenziell auf fehlende Blätter in der Lagenmitte der Vorlage zurückzuführen oder Texterweiterungen mit verfügbarem Freiraum am Ende einer Lage erklärbar.

Die Perzeption der Zusammenhänge zwischen Lagen- und Blattstrukturen einerseits und dem Text andererseits ist für den Benutzer herkömmlicher Ausgaben kaum möglich, selbst wenn einer Edition eine Beschreibung des texttragenden Artefakts beigelegt ist. Bei der Textlektüre sind Informationen dieser Art in herkömmlichen Ausgaben nicht direkt verfügbar. Selbst dort, wo bisher versucht wurde, Lagenstrukturen als Begleitfunktion eines digitalen Faksimiles zu visualisieren (`<Canterbury Tales Project>`, `<Parzival-Projekt>`),

wurden lediglich statische Lagenskizzen mit Einzelseiten verknüpft, sodass weder ein analytischer Zugriff noch eine dynamische Navigation und Visualisierung möglich waren.

Unser Ansatz, der im Zusammenhang mit der zur Zeit entstehenden Plattform ›Welscher Gast digital‹ (einem Kooperationsprojekt des Sonderforschungsbereichs 933 ›Materiale Textkulturen‹ und der Universitätsbibliothek Heidelberg) entwickelt wird, setzt bei der TEI-konformen Kodierung der physischen Lagen- und Blattstrukturen mittelalterlicher Handschriften des ›Welschen Gastes‹ an, die direkt im Code der Texttranskription notiert werden. Dadurch werden Abfragen möglich über die Zusammenhänge zwischen hierarchischen Strukturen des Werkes (Bücher, Kapitel, Verspaare, Verse) und materiellen Strukturen des texttragenden Artefakts. Bei der Kodierung arbeiten wir mit mehreren Typen der <surfaceGrp>-Elemente (*binding, gathering, bifolium, leaf*), die durch ihre Schachtelung die physische Zusammensetzung der Kodizes abbilden. Eventuelle Defekte (fehlende Blätter) und Ergänzungen (eingeklebte oder eingenähte Zusatzblätter) werden an entsprechenden Elementen durch Attribute und fehlende bzw. zusätzliche Knoten direkt realisiert, womit die Lagenzusammensetzung präzise beschrieben ist.

Auf der Basis dieser Kodierung (und einer aus Performanzgründen daraus generierten relationalen Struktur) entwickeln wir eine visuelle SVG-Schnittstelle, die dem Benutzer des digitalen Faksimiles eine Navigation durch die physischen Kodexstrukturen und einen davon ausgehenden Zugang zu Text und Bild ermöglicht. Der Benutzer kann dadurch eine konkrete Lage ansteuern und darin schematisch blättern.

An den Eckpunkten der physischen Struktur (Seitenumbrüche, Blattwechsel, Lagengrenzen) wird zudem das Zusammenfallen oder die Überlappung mit feinkörnigen hierarchischen Strukturen des Werkes (Vers- und Doppelpersengrenzen) durch farbig differenzierte Symbole angezeigt. Schließlich visualisieren parallel mit der Lagenanzeige verlaufende Farbleisten Übereinstimmungen und Unterschiede der materiellen Einheiten des Buches und der ideellen Makrostrukturen des Werkes.

Anstelle einer separaten Beschreibung befolgt unser Ansatz die Maxime einer in die digitale Edition integrierten Veranschaulichung. Damit stehen nicht nur die abgelegten Daten der wissenschaftlichen Analyse zur Verfügung, sondern die Visualisierungen an sich erschließen neue Perspektiven auf das physische ›Gewordensein‹ und die zugrundeliegende Planung der Kodizes. Auf diese Weise vermitteln die Darstellungen nicht nur die Datenbasis, sondern sind – ganz im Sinne der digitalen Geisteswissenschaften – ihrerseits eigenständige Impulsgeber für neue Interpretationen, die wiederum zum Ausgangspunkt neuer Fragestellungen werden können.

Sonic materialization of linguistic data

The problem of sonification

Kramer Gregory (1994) in his book *“Auditory Display: Sonification, Audification, and Auditory Interfaces”* defines sonification as “use of non-speech audio to convey information or perceptualize data”.

In our digital age we can store, edit and examine almost all qualities and quantities as data. Sound itself can be considered as a pure stream of information able to be modulated, transformed and analyzed in a lot of different ways.

The success of sonification occurs when the sound reveals one or more qualities of data or data reveals one or more qualities of sound. Thus, this kind of materialization of data is an interdisciplinary act which involves both the proper analysis of data and the structure of sound.

While technology provides us with a wide variety of tools, the core of the problem still exists. As this kind of interdisciplinary knowledge is hard to be combined, there aren't enough available tools which help artists to escape from an arbitrary mapping of data to sound qualities. This leads to arbitrary results both for the artist and the listener as the sonification process doesn't take advantage of neither the auditory perception properties nor sound's advantages in temporal, amplitude and frequency resolution. As a result, in most cases, sonification fails its purpose which “is to encode and convey information about an entire data set or relevant aspects of the data set.” (The Sonification Handbook 2011)

Sound and linguistics

“Sonic Materialization of Linguistic Data“ is a series of work and a research project aiming to provide sound artists with the tools for the proper linguistic analysis of the mined data.

In our age of constant connectivity, social media - and especially the twitter text-based platform- can be considered as a monitor corpus which evolves perpetually and it is in a process of constant change. In order to create new structures and transform this chaotic

stream of data into new material - in our case sound, it needs to be organized according to its different kind of properties- here its linguistic aspects. With our work "Sonic Materialization of Linguistic Data" we provide different software modules that can perform real time linguistic analysis of data and output the result for sonification purposes.

Our software consists of different kind of modules, from which the user can choose only one or a combination of more. Here we present the *Stress Module*. The program enables the user to aggregate data from different hashtag [#] feeds on twitter in real time. The incoming data is being processed according to their linguistic features and in particular stress. The algorithm performs a series of tasks and extracts the stressed syllables of the aforementioned data. The output is a phonetic transcription code which represents each phoneme of the input twitter feed. The encoded outputted list of data also includes suggestions for the sonic mapping that occurs from data's linguistic features and the sound's nature. For instance, the strong syllables are a numerical output which represents a longer sound event (time envelope), whereas the weaker syllables are a numerical representation of a briefer sound event. Similar kind of optional mapping can also affect other sound features such as pitch, timbre, ADSR envelopes, modulation etc.

Stress, which can be considered as a prosodic feature, manifests itself in the speech stream in several ways. Stress patterns seem to be highly language dependent, considering that there is a dichotomy between stress timed and syllable timed languages. In stress timed languages primary stress occurs at regular intervals, regardless of the number of unstressed syllables in between, whereas in syllable timed languages syllables tend to be equal in duration and therefore are inclined to follow each other at regular intervals of time. According Halliday(1985: 272), "salient syllables occur in stress timed languages at regular intervals". Strong syllables bear primary or secondary stress and contain full vowels, whereas weak syllables are unstressed and contain short, central vowels.

Particularly in English, which is a stress language, speech rhythm has a characteristic pattern which is expressed in the opposition of strong versus weak syllables. Stressed syllables in English are louder, but they also tend to be longer and have a higher pitch. Despite the fact that stress can be also influenced by pragmatic factors such as emphasis, our project aims to capture the natural stress pattern of English in order to

extract meaning from sound patterns too, as they will be delineated by the phonetic structure of natural language.

Presentation

For the presentation of the project we are proposing a poster with the description of how exactly the software works and what its aim is. We also would like to include a pair of headphones and a small screen (or projector) in order to have the data analysis and the sonification process in real time for the audience to experience.

References

Kramer, Gregory 1994. Auditory Display: Sonification, Audification, and Auditory Interfaces. Santa Fe Institute Studies in the Sciences of Complexity, Proceedings Vol. XVIII. Addison Wesley, Reading, Mass.

Halliday, M. A .K. 1985. An Introduction to Functional Grammar. London: Arnold.

The Sonification Handbook. Edited by Thomas Hermann, Andy Hunt, John G. Neuhoff (Eds.). Berlin: Logos Verlag Berlin 2011

Zur OCR frühneuzeitlicher Drucke am Beispiel des RIDGES-Korpus von Kräutertexten

Uwe Springmann (LMU München und Humboldt-Universität zu Berlin)
& Anke Lüdeling (Humboldt-Universität zu Berlin)

Unser Vortrag stellt eine Methode zur optischen Zeichenerkennung (OCR) von frühen Drucken vor, die deutlich bessere Resultate zeigt als vorherige Methoden. Mithilfe des Verfahrens können leichter und schneller Korpora mit frühen Texten erstellt werden, die dann nur noch nachkorrigiert werden müssen. Mit dem Aufbau solcher Korpora aus frühneuzeitlichen Drucken werden Basisressourcen für alle darauf aufbauenden Forschungen sprachlicher, historischer und kulturgeschichtlicher Art in den Digital Humanities bereitgestellt. Wir exemplifizieren unsere Methode mit Daten aus dem RIDGES-Korpus¹, einem diachronen Korpus, das deutschsprachige Kräutertexte enthält, die zwischen 1487 und 1870 entstanden sind.

Mangels maschineller Unterstützung ist die Erstellung eines solchen Korpus aufwändig und vom Finden korpusrelevanter gut lesbarer Vorlagen über die Einweisung von Hilfskräften in die Transkription ungewohnter Zeichen und paläographischer Konventionen und einer für die Korrektur der Transkription notwendigen breiten Sprach- und Sachkenntnis geprägt. Eine historische Orthographie und der unvermittelte Wechsel von deutschem Fraktur-Text zu lateinischen Zitaten in Antiqua sowie griechischen Wörtern erschweren die Erstellung der Transkription zusätzlich. Insbesondere die frühen Drucke (Wiegendrucke, aber auch noch Drucke aus dem 17. Jahrhundert) sind hier schwierig.

Der Traum von einer automatischen Unterstützung bei der Konvertierung früher Drucke durch allgemein zugängliche Methoden einer OCR, die ein entsprechendes Training der Erkennungsroutinen auf die verwendeten Schriften sowie die Eigentümlichkeiten des Druckbildes voraussetzen, ließ sich bisher angesichts proprietärer, einem umfangreichen Training für Außenstehende nicht zugänglicher Industrieprodukte (z.B. Abby Finereader²) bzw. zwar quelloffener und grundsätzlich trainierbarer, aber an der gestellten Aufgabe scheiternder Software (z.B. Tesseract³) nicht verwirklichen. Neben diesen grundsätzlichen Mängeln stand einem solchen Ansatz bisher auch der Umstand entgegen, dass ein Training eine systematisch erstellte diplomatische, d.h. am Druckbild orientierte und nicht-normalisierende Transkription von Texten voraussetzt.

Im Jahr 2013 wurden die bei Mustererkennungsaufgaben sehr erfolgreichen rekurrenten neuronalen Netzwerke mit langem Kurzzeitgedächtnis (LSTM: long short-term memory; Hochreiter & Schmidhuber 1997) durch Thomas Breuel erstmals in die OCR eingeführt und in das quelloffene, schon länger bestehende System OCropus (Version 0.7)⁴ integriert (Breuel et al. 2013). Das RIDGES-Korpus enthält Textausschnitte aus vielen Kräuterbüchern. Diese Ausschnitte (meist ca. 30 Textseiten) wurden eng diplomatisch transkribiert.⁵ Das Training dieses Systems mit Hilfe der diplomatischen Transkription ("ground truth") zeigt Ergebnisse, die bei jedem der vorliegenden Texte eine Rate korrekt erkannter Zeichen (einschließlich Ligaturen, Diakritika

1 Ridges steht für Register in Diachronic German Science; Ziel des Ridges-Projekts ist die qualitative und quantitative Analyse der Entstehung eines deutschsprachigen wissenschaftlichen Registers. Dazu gibt es viel Literatur (so z. B. Klein 2011 oder Habermann 2003), die sich bisher aber auf nicht digital vorliegende Texte stützen musste und daher kaum für statistische Registeranalysen ausgewertet werden konnte. Das Korpus ist unter der CC-BY-Lizenz verfügbar unter http://korpling.german.hu-berlin.de/ridges/index_de.html. Das Korpus ist tief annotiert und wächst ständig.

2 <http://www.abbyy.de/>

3 <http://code.google.com/p/tesseract-ocr/>

4 <http://www.ocropus.com>

und Leerzeichen) von über 96% selbst ohne Verwendung von Sprachmodellen und Nachkorrekturen ergibt, während bisherige Versuche mit kommerziell erhältlicher Software bzw. Tesseract, an denen ein Autor seit Jahren beteiligt ist, kaum an die Grenze von 90% heranreichen (Springmann et al. 2013).⁶

**vbergſchlagēpflaſters weiſ wirt/
verhütet ſys für dē kaltē brandt/
beylet darzū mercklich bald zūſa-
men gleichſam der walwurzen/
laſt nicht bald die zūuellige hitz
vberhand nemmen.**

Ariſtolochia rotunda.

**Ariſtolochia wachſet auff hohen
wyſen mit einer runden wurzen/
änlich cyclamini wurzel/ auffge-
nomē dz diſe iſt inwendig gālb/
eines bitteren ſtarcken geruchs/
auff jhren wachſend viel ſubteile
zāſerlin/ welche ſich oben als riet-
lin oder zincklein herfür thünd/
die habend kleine ſchier anzūſähē
als Ebheūw bletter/ bringend im
ſommer gwonlich herfür bleich-
gelbe blümē / Diſ ſchōn gewächs
hab ich nie friſch / das iſt grien o-
der lebend in teüdtſchem land ge-
ſehen / Es vergleichet ſich weder
am ſtengel/ kraut/ noch in zeit ſei-**

vbergſchlagēpflaſters weiſ wirt /
verhütet ſys für dē kaltē brandt /
heylet darzū mercklich bald zūſa-
men gleichſam der walturtzen /
laſt nicht bald die zūuellige hitz
vberhand nemmen.

Ariſtolochia rotunda.

Ariſtolochia wachſet auff hohen
wyſen mit einer runden wurzen /
änlich cyclamini wurzel / auffge-
nomē dz diſe iſt inwendig gālb /
eines bitteren ſtarcken geruchs /
auff jhren wachſend viel ſubteile
zāſerlin / welche ſich oben als riet-
lin oder zincklein herfür thünd /
die habend kleine ſchier anzūſähē
als Ebheūw bletter / bringend im
ſommer gwonlich herfür bleich-
gelbe blümē / Diſ ſchōn gewächs
hab ich nie friſch / das iſt grien o-
der lebend in teüdtſchem land ge-
ſehen / Es vergleichet ſich weder
am ſtengel / kraut / noch in zeit ſei-

Adam von Bodenstein (1557): Wie sich meniglich Unkorrigierter OCR-Output einer vorher nicht gesehenen Seite nach Training auf 49.000 zufällig ausgewählten Textzeilen (Bild + zugeordnete ground truth) aus einer Trainingsmenge von 34 diplomatisch transkribierten Seiten. Die OCR zeigt 7 verbleibende Fehler auf dieser Seite (das entspricht der durchschnittlichen Zeichenerkennungsrate auf einer Testmenge von Seiten von 99,0%).

Der Grund für diese hochgenaue Erkennungsrate liegt darin, dass OCRopus im Gegensatz zu bisherigen Methoden keine Erkennung auf Zeichenbasis über ein Template-Matching-Verfahren durchführt, bei dem ein errechnetes „mittleres“ Zeichen (das Template) auf Übereinstimmung mit einem zu erkennenden Zeichen überprüft wird, sondern jede Druckzeile durch Zerlegung in bis zu 1000 vertikale

⁵ Die Transkription wurde von Studierenden in verschiedenen Seminaren begonnen und später korrigiert. Daneben gibt es zwei Normalisierungsebenen und verschiedene Annotationsebenen.

⁶ Lediglich für die kommerzielle Software B.I.T. Alpha wurden ähnlich gute Ergebnisse für Drucke des 16. und 17. Jahrhunderts berichtet, wobei die erreichbare Genauigkeit von einem für Außenstehende kaum nachzuvollziehenden In-House-Training des kommerziellen Anbieter abzuhängen scheint (Stäcker in Federbusch et al. 2013).

Streifen schneidet, so dass jeder Buchstabe und jeder Wortzwischenraum in bis zu 30 Streifen zerlegt wird. Für jeden Streifen werden im Laufe des Trainings über den Vergleich von gedruckter Zeile mit ihrer zugeordneten Transkription die Parameter des neuronalen Netzes so eingestellt, dass mit hoher Wahrscheinlichkeit der richtige Buchstabe ausgegeben wird. Die Übersegmentierung der Buchstaben führt zu einer höheren Auflösung bei der Erkennung, so dass auch zwischen ähnlichen Zeichen wie langem s (f) und f problemlos unterschieden werden kann.

Die Aussicht, dass sich nunmehr jeder Interessierte Texte in hoher Genauigkeit in elektronischer Form verschaffen kann, selbst wenn die zugrundeliegenden Drucke aus früheren Jahrhunderten stammen, wird anhand unserer Erfahrungen mit dem RIDGES-Korpus hinsichtlich ihrer Voraussetzungen und des damit verbundenen Aufwandes kritisch beleuchtet. Dabei werden sowohl die Rolle des Trainings als auch der Nachkorrektur sowie die Stellung der OCR im gesamten Prozess der Korpuserstellung diskutiert. Die verwendeten Werkzeuge sowie Trainings- und Testdaten samt einer Anleitung zur Nutzung des Systems werden unter einer Open-Source-Lizenz veröffentlicht und stehen der Allgemeinheit in Kürze zur Verfügung.

Referenzen

Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A., & Shafait, F. (2013). High-performance OCR for printed English and Fraktur using LSTM networks. In *Document Analysis and Recognition (ICDAR), 2013*, 683–687.

Federbusch, M., Polzin, C., & Stäcker, T. (2013). *Volltext via OCR - Möglichkeiten und Grenzen: Testszenerien zu den Funeralschriften der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz. Erfahrungsbericht aus dem Projekt "Helmstedter Drucke Online" der Herzog August Bibliothek Wolfenbüttel/von Thomas Stäcker*. Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Berlin.

Habermann, M. (2003). Der Sprachenwechsel und seine Folgen. Zur Wissensvermittlung in lateinischen und deutschen Kräuterbüchern des 16. Jahrhunderts. In: *Sprachwissenschaft* 28, 325–354.

Klein, W.-P. (2011). Die deutsche Sprache in der Gelehrsamkeit der frühen Neuzeit. Von der *lingua barbarica* zur *Hauptsprache*. In: Jaumann, Herbert (Hg.) *Diskurse der Gelehrtenkultur in der Frühen Neuzeit. Ein Handbuch*. de Gruyter, Berlin/New York, 465–516

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., & Fink, F. (2014). OCR of historical printings of Latin texts: problems, prospects, progress. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 71–75.

StreetartFinder – Eine Datenbank zur Dokumentation von Kunst im urbanen Raum

1. Einleitung

Streetart ist ein Sammelbegriff für Graffitis, schablonen- oder handgezeichnete Bilder, Poster, Aufkleber aber auch Installationen und Skulpturen im öffentlichen Raum (Reinecke, 2012, S. 17). Eine Vielzahl bestehender Publikationen zeigt, dass Streetart auch als wissenschaftliches Forschungsobjekt zunehmend an Relevanz gewinnt (vgl. etwa Klitzke & Schmidt, 2009; Philipps & Barlösius, 2014; Reinecke 2012; Waclawek, 2012; u.v.a.). Mit dem *StreetartFinder*¹ wurde ein Tool geschaffen, das erlaubt, diese Kunstwerke im urbanen Raum in digitaler Form zu dokumentieren, und so eine Datenbank für weitere Forschung in diesem Feld zur Verfügung zu stellen.

2. Konzeption und wesentliche Funktionen des StreetartFinder

StreetartFinder wurde mit gängigen Web-Technologien umgesetzt, und steht sowohl als *Desktop*- als auch als *Mobile*-Variante zur Verfügung. Nutzer können Fotos von Streetart-Objekten auf die Webseite laden, und dabei Metadaten wie „Name des Uploaders“, „Tags / Schlagworte“ sowie einen optionalen „Beschreibungstext“ angeben. Zusätzlich sind die Uploader angehalten, ihr jeweiliges Objekt zu klassifizieren, wobei derzeit folgende Optionen zur Auswahl stehen: *Graffiti*, *Stencil*, *Painting*, *Paste-Up*, *Installation*, *Sonstiges*. Zuletzt können die Nutzer optional den Standort der Streetart durch Markierung in einem *GoogleMaps*-Ausschnitt vornehmen.

Streetart kann gefiltert nach Städten, Kategorie, Bewertung oder Anzahl der Views dargestellt werden (vgl. Abb. 1). Die Besucher der Seite können die bestehenden Streetart-Fotos bewerten oder aber ein Objekt als „nicht länger vorhanden“ melden.

¹ <http://streetartfinder.de/>, zuletzt aufgerufen am 23.10.2014

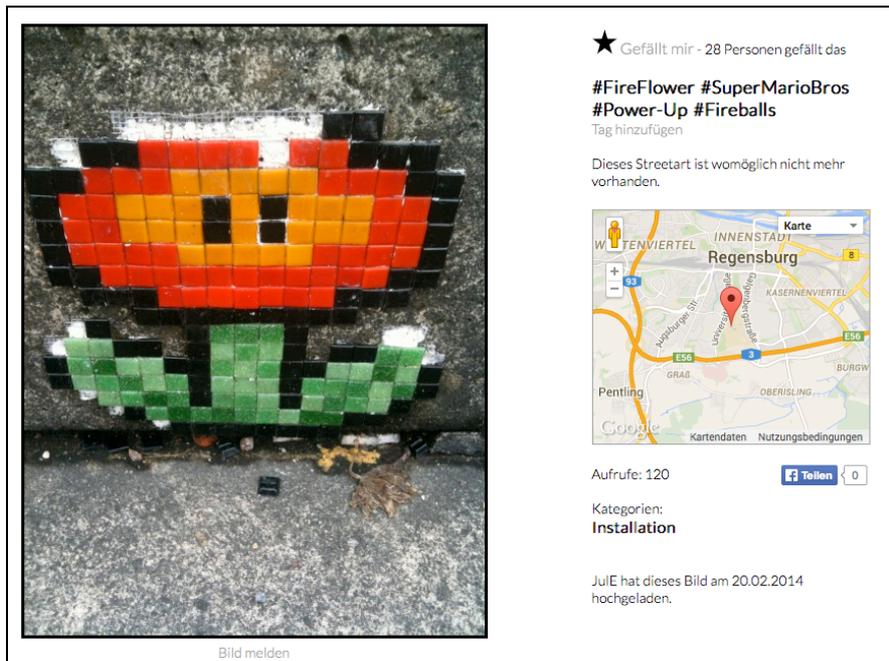


Abbildung 1: Darstellung eines Streetart-Objekts mit verschiedenen Metadaten auf der Webseite.

Eine wesentliche Funktion stellt außerdem die Visualisierung verschiedener Streetart-Objekte auf einer interaktiven Karte dar, die mit Hilfe der *GoogleMaps API*² realisiert wurde (vgl. Abb. 2). Auf dieser Karte kann etwa dargestellt werden wo sich welche Art von Streetart befindet.

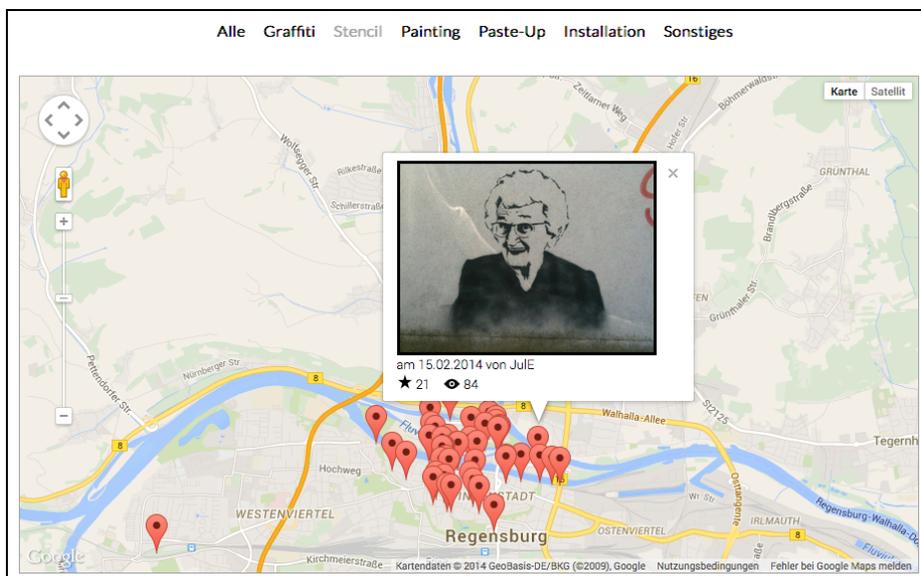


Abbildung 2: Interaktive GoogleMaps-Visualisierung der Streetart-Objekte.

² <https://developers.google.com/maps/>, zuletzt aufgerufen am 23.10.2014

4. Fazit

Die bisherigen Nutzerzahlen zeigen, dass *StreetartFinder* von den Benutzern als Tool zur Dokumentation von Kunst im urbanen Raum gut angenommen wird. Es entsteht auf diese Weise eine einzigartige Datenbank, in der neben Fotografien der jeweiligen Streetart auch Metadaten mitgespeichert werden, die verschiedene soziologische, kultur- und medienwissenschaftliche Fragestellungen erlauben, z.B.:

- Welcher Typ von Streetart kommt am häufigsten vor?
- Gibt es im Laufe der Zeit Trends für bestimmte Typen bzw. gibt es Ballungsgebiete, in denen vor allem ein bestimmter Typ von Streetart vorherrscht?
- Wie lange ist die durchschnittliche Lebensdauer von Streetart, und gibt es einen Zusammenhang mit dem Ort / Typ?
- Was sind die Hauptfunktionen von Streetart?

Neben Überlegungen zur weiteren Verbreitung des Tools, vor allem auch in anderen Städten, planen wir zusätzlich einen Web-Zugang zu allen relevanten Metadaten für interessierte Wissenschaftler.

5. Literaturverzeichnis

Klitzke, K. & Schmidt, C. (2009). *Street Art: Legenden zur Straße*. Berlin: Archiv der Jugendkulturen.

Philipps, A. & Barlösius, E. (2014). Zur Sichtbarkeit von Street Art in Flickr. Methodische Reflexionen zur Zusammenarbeit von Soziologie und Informatik. In *Abstracts of the DhD 2014*, Passau.

Reinecke, J. (2012). *Street-Art: eine Subkultur zwischen Kunst und Kommerz*. Bielefeld: Transcript Verlag.

Waclawek, A. (2012). *Graffiti und Street Art*. Berlin: Deutscher Kunstverlag.

Virtuelle Rekonstruktion des Regensburger Ballhauses

1. Projektkontext und wesentliche Ziele

Im Rahmen einer Vortragsreihe zum 350-jährigen Reichstagsjubiläum in der Stadt Regensburg wurde in Ergänzung zum Thema „Das Jahrhundert des Dramas und der Komödien: Blüte des Regensburger Theaterlebens“¹ eine virtuelle 3D-Rekonstruktion des heute nicht mehr vorhandenen Regensburger Ballhauses am Ägidienplatz erstellt. Die 3D-Rekonstruktion stellt einerseits das Innenleben des Ballhauses dar und liefert andererseits textuelle Informationen zu interessanten Objekten. Die Rekonstruktion kann mit Hilfe der Virtual Reality-Brille *Oculus Rift* interaktiv exploriert werden. Das Projekt ist damit im Kontext der Museumspädagogik anzusiedeln (vgl. Flügel 2009; Wagner 2007; Waidacher & Raffler 2005).

Umfangreiche Informationen zur Geschichte des Regensburger Ballhauses am Ägidienplatz finden sich in Meixner (2008): Die Baugeschichte des Ballhauses beginnt bereits im Jahre 1603, als zunächst hölzerner Bau, der vornehmlich für Sportereignisse genutzt wurde. Dieses Gebäude wurde schließlich im Jahre 1736 durch einen Neubau ersetzt, der dann auch stärker für Theateraufführungen genutzt wurde. In seiner Hochzeit war das Ballhaus am Ägidienplatz das kulturelle Zentrum des Immerwährenden Reichstags in Regensburg. Durch die Eröffnung des Theaters am Bismarckplatz im Jahre 1804 verlor das Ballhaus langsam an Bedeutung. In der Folge verfällt das Gebäude zunehmend und wird schließlich im Jahre 1922 abgerissen.

Hauptziele des Projekts

- Rekonstruktion des Innenraums des Ballhauses (1736-1922) mit der Präsentation einer barocken Kulissenbühne
- Interaktion durch *Virtual Reality*-Umsetzung statt statische Präsentation eines 3D-Modells
- Umsetzung einer zusätzlichen pädagogische Komponente durch das Augmentieren weiterführender Information über das Ballhaus im virtuellen 3D-Raum

2. Unvollständigkeit der Quellenlage als wesentliche Herausforderung

Die exakte Gestaltung des Innenraums ist sehr schwer zu rekonstruieren, da nur wenige Quellen aus dieser Zeit überliefert wurden. Soweit Quellen vorliegen, sind diese zumeist Skizzen von Zeitzeugen und Dokumente aus dem Hofarchiv Thurn und Taxis (vgl. Abb. 1).

¹ Referentin: Hannah Ripperger; weitere Informationen zum Vortrag im Programmheft zur Vortragsreihe (S. 17), online verfügbar unter: <https://www.regensburg.de/sixcms/media.php/121/der-reichstag-in-45-minuten.pdf>, zuletzt abgerufen am 27.10.2014.

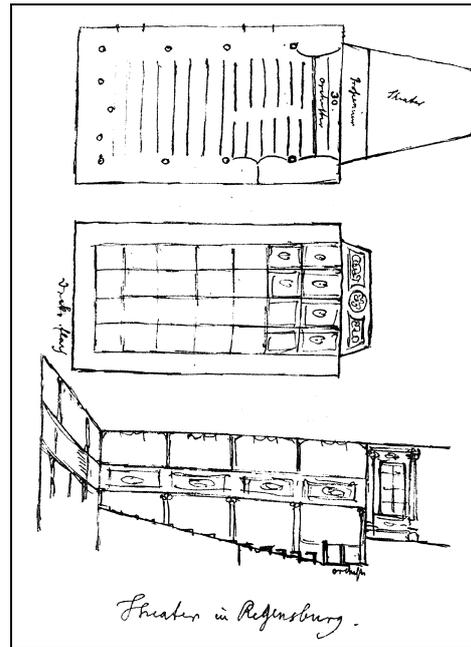


Abbildung 1: Skizzen von Friedrich Gilly, 1798 (Bildquelle: Meixner, 2008, S. 126).

Weitere Herausforderungen ergeben sich durch historische Maßeinheiten (z.B. „Regensburger Schuh“ statt Meter), oder durch Skizzen ohne Maßstab und Maßangaben. Zudem gibt es oftmals keine Abgrenzung zwischen verschiedenen Bauphasen des Hauses. Um diese unvollständigen oder fehlenden Informationen zu ergänzen, wurden schließlich Vergleiche zu anderen Theaterräumen in Deutschland angestellt (etwa Gotha und Passau), und allgemeine Stilmerkmale aus der Kunstgeschichte umgesetzt (vgl. Meixner, S.128 f). Diese heterogene Quellenlage ist in Abbildung 2 zusammengefasst dargestellt:

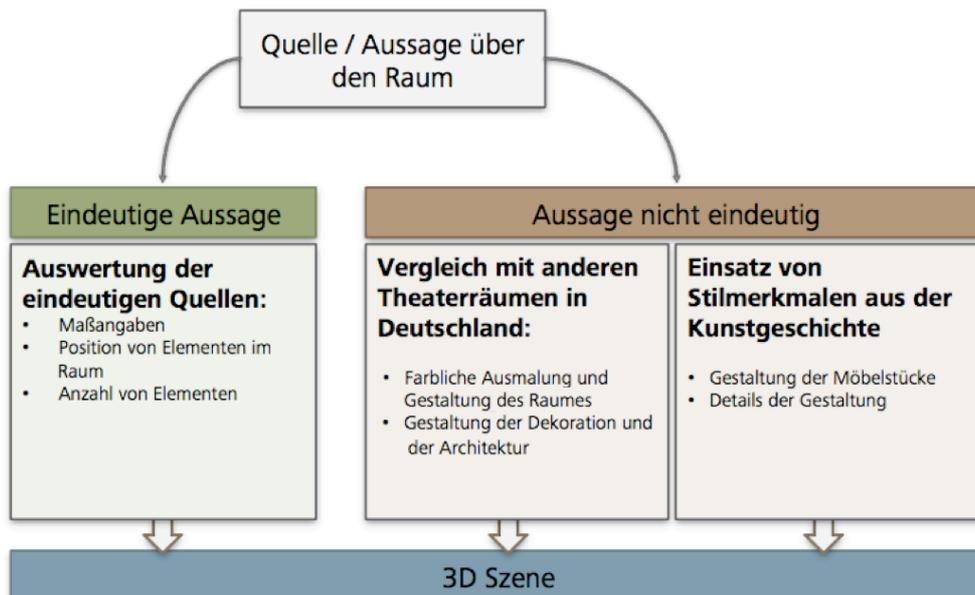


Abbildung 2: Überblick zur heterogenen Quellenlage, die als Grundlage für die Rekonstruktion verwendet wurde.

3. Technische Umsetzung

Die Rekonstruktion wurde mit Hilfe des 3D-Modellierungstools *Blender*² umgesetzt. Zu Beginn wurde die Geometrie des Theatersaals sowie dessen Möblierung modelliert. Zusätzlich wurde mit Hilfe der Bildbearbeitungssoftware Photoshop und einigen Referenzbildern die farbliche Gestaltung des Raums nachgebildet (vgl. Abb. 3).



Abbildung 3: Ausschnitt aus der virtuellen Rekonstruktion des Regensburger Ballhauses am Ägidienplatz.

Nach der Rekonstruktion des Raums folgte der Export in die Game-Engine *Unity3D*³. Dort wurde die Geometrie mit Farbinformationen und mit Oberflächenstrukturen ausgestattet. Danach wurden alle weiteren Interaktionsmöglichkeiten implementiert. Dabei wurde auf das *Oculus Rift SDK*⁴ zurückgegriffen: Zwei Kameras rendern die Szene und verkrümmen das gerenderte Bild, um es gemäß der Linsenkrümmung im *Head Mounted Display* (HMD) korrekt anzeigen zu können (vgl. Abb. 4).

² <http://www.blender.org/>, zuletzt abgerufen am 27.10.2014

³ <http://unity3d.com/>, zuletzt abgerufen am 27.10.2014

⁴ <http://www.oculus.com/>, zuletzt abgerufen am 27.10.2014



Abbildung 4: 3D-Szene aus Perspektive der Virtual-Reality-Brille *Oculus Rift*.

Mithilfe eines einfachen Game-Controllers kann sich der Nutzer im Raum bewegen. Ferner kann über die Bewegung mit dem Kopf die Rotation der Kamera bestimmt werden. Über den Mittelpunkt des Bildschirms und entsprechendes *Raycasting*⁵ in den Raum wird überprüft, welches Objekt der Nutzer gerade anschaut. Je nachdem können verschiedene Informationen über die Elemente im Raum, etwa die Kulissenbühne (vgl. Abb. 5), mit einem Tastendruck über den Game-Controller abgerufen werden.

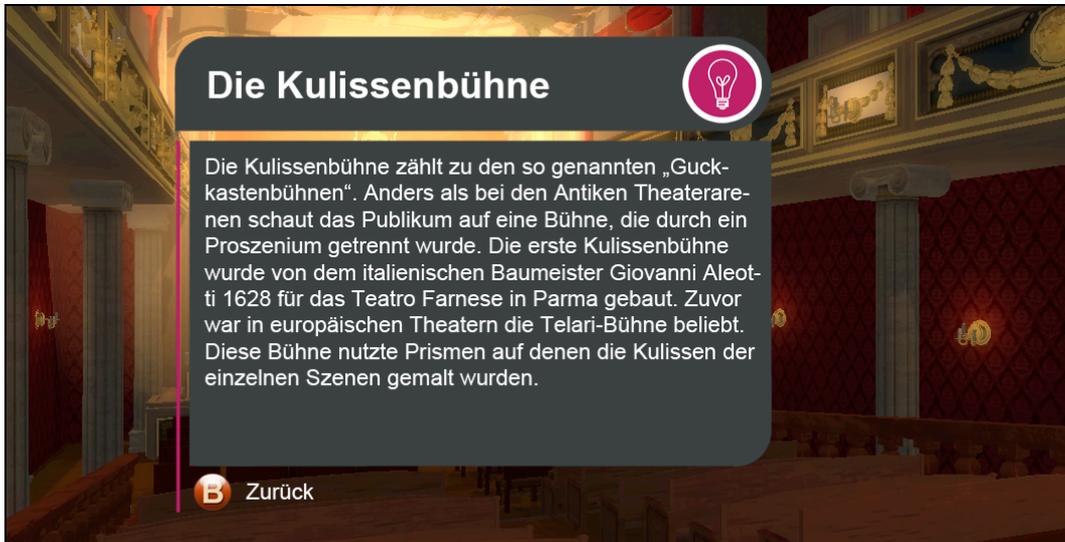


Abbildung 5: Informationsanzeige zur Funktionsweise der Kulissenbühne.

⁵ *Raycasting* ist ein Begriff aus der Computergrafik. Vereinfacht gesagt „tastet“ dabei ein virtueller Strahl den dreidimensionalen Raum nach Objekten ab, die dann bei Bedarf aktiviert werden können.

4. Demonstration

Ein Demo-Video der virtuellen Rekonstruktion ist verfügbar unter:

- <http://dhregensburg.wordpress.com/2014/07/25/virtuelle-rekonstruktion-regensburger-ballhaus/>

Im Falle der Annahme des Abstracts ist geplant, die Rekonstruktion im Rahmen der Poster-Präsentation live mithilfe einer Virtual-Reality-Brille (*Oculus Rift*) zu demonstrieren.

5. Literaturverzeichnis

Flügel, K. (2009). Einführung in die Museologie. 2., überarb. Aufl. Darmstadt: WBG.

Meixner, C. (2008). Musiktheater in Regensburg im Zeitalter des Immerwährenden Reichstages. Sinzig: Studio Verlag.

Wagner, E. (2007). Museum, Schule, Bildung. Aktuelle Diskurse, innovative Modelle, erprobte Methoden. München: kopaed.

Waidacher, F. & Raffler, M. (2005). Museologie - knapp gefasst. 1. Aufl. Stuttgart: UTB.

Statistisch gestützte Visualisierung von Informationsgliederungen in den Bundestagsreden

Dr. Zakharia Pourtskhvanidze, pourtskhvanidze@em.uni-frankfurt.de
Institut für Empirische Sprachwissenschaft. Goethe-Universität Frankfurt/M

Der handlungsbezogene Aspekt der Sprache wird besonders deutlich in den öffentlichen Auftritten der Politiker. Neben der für die gesprochene Sprache spezifischen Verwendung von lexikalischen Mitteln erscheinen in den Reden zuhörerorientierte Einsetzung von syntaktischen Konstruktionen (Z.B. Anrede, Rhetorische Fragen) und auf die Interaktion abgestimmte Realisierung von pragmatischen kommunikativen Strategien (Z.B. Gesichtserhaltung).

Der linguistisch interpretierte Begriff *Hervorhebung* bezüglich der Gestaltung des informationellen Gehalts einer Äußerung eignet sich besonders als ein Ausgangspunkt für die Verbildlichung (Visualisierung) besonders prominenter Informationseinheit im Vergleich zur informationell neutralen Einheit eines gesprochenen Diskurses.

Die **empirische Basis** der auf dem Poster beschriebenen Analyse stellt das Plenarprotokoll (Stenographischer Bericht) der 2. Sitzung des Deutschen Bundestages am 18. November 2013 zum Thema *die Abhöraktivitäten der NSA und die Auswirkungen auf Deutschland und die transatlantische Beziehungen* dar (ca. 20.000 Token bei 17 Sprechern aus 5 Parteien).

Aufgrund der linguistischen Analyse von Texten wird eine **Ranking-Tabelle** der grammatischen Instrumente der Informationsgliederung erzeugt. In der Tabelle bekommen die einzelnen Instrumente (Partikeln, Vorfeldbesetzung, Spalt-Satzstruktur, Satzarten, Referentielle Bezüge (Anapher bzw. Katapher) etc.) in den Zahlen ausgedruckte Werte (-15 min. ... 0 ... 75 max.). Die fokussierende (rhematisierende) Instrumente bekommen tendenziell hohe Werte (daher stärker hervorgehoben), wogegen die topikalisierte (thematisierende) Instrumente bekommen tendenziell niedrige Werte (daher flachere Visualisierung).

Der *erste Schritt* des Visualisierungsvorgangs sieht die automatische **Ersetzung** des jeden in der Ranking-Tabelle notierten Tokens resp. Skopus im Protokoll-Text mit der entsprechenden Zahl, während der Rest von Tokens im Text automatisch mit dem Wert „0“ ersetzt wird. Die Texte werden in CSV-Dateiformat abgebildet und damit für unterschiedliche Statistik-Analysen endgültig aufbereitet.

Die Visualisierung der Daten erfolgt im *zweiten Schritt* durch eine freie Programmiersprache für statistisches Rechnen und statistische Grafiken **R**.

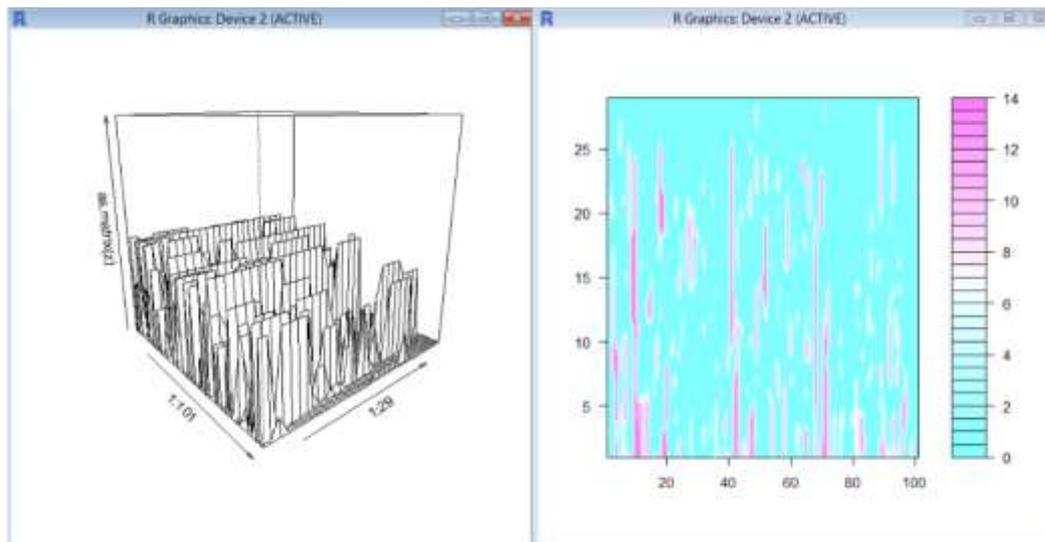


Abb. 1. Visualisierung der Rede des Abgeordneten G. Gysi (Die Linke) mit R-Funktionen: *persp* (links) und *filled.contour* (rechts)

Im Kontext der Auffassung *Make your data tell a story* eröffnet die Visualisierung von Informationsgliederungen die Möglichkeiten die generierten Bilder auf die Zeitschiene zu ordnen und zwar limitiert auf eine Person oder einem Themas.

- Gibt es mögliche Kausalitätstendenzen zwischen den „Informationslandschaften“ der Reden und dem Thema desselben?
- Lässt sich ein Dominanz-Muster für die bestimmten sprachlichen Instrumente der Informationsgliederung abhängig vom Thema etablieren?
- Gibt es individuelle Muster von informationellen Hervorhebungen?
- Lässt sich die Entwicklung des Redestils eines Sprechers in den Visualisierungen von Informationseinheiten über die Zeit dokumentieren?

Eine disziplinübergreifende Analyse mit den Bezügen auf die Gender- und Meinungsforschungen ist denkbar.

Vorgesehen ist die Erweiterung der empirischen Daten in Richtung der Vielfalt der Genres und die Einsetzung von alternativen Visualisierungssoftware (z.B. RStudio).

Das weitere Forschungsvorhaben wird gegenwärtig an der Goethe-Universität für die Aufbau einer integrativen Kooperationsplattform zwischen den Fachbereichen „Sprach- und Kulturwissenschaften“ und „Informatik und Mathematik“ bedacht.

Vom Luftbild zum 3D-Modell

zum Einsatz von unbemannten Luftfahrzeugen in der Archäologie des Vorderen Orients

Benjamin Glissmann, Jason Herrmann¹, Matthias Lang²

¹ Institut für die Kulturen des Alten Orients der Universität Tübingen, ² eScience-Center der Universität Tübingen

Unbemannte Luftfahrzeuge – meist als Drohnen bezeichnet – entwickeln sich immer mehr zu einem wichtigen Dokumentationswerkzeug in der archäologischen Feldforschung. Waren die Geräte noch vor wenigen Jahren äußerst kostspielig sowie schwierig zu fliegen und zu warten, lassen sich heute bereits für niedrige dreistellige Beträge Drohnen erwerben, deren Betrieb auch durch einen Laien schnell zu erlernen ist.

In einem Großteil der Projekte dienen die UAVs (unmanned aerial vehicle) meist zur Aufnahme von Luftbildern und Filmen zu reinen Visualisierungs- und Präsentationszwecken, eine Integration der Geräte in den eigentlichen Dokumentations- und Forschungsprozess findet meist nicht statt.

Integration von unbemannten Luftfahrzeugen in die archäologische Dokumentation

In diesem Beitrag soll diskutiert werden, wie diese Systeme sinnvoll in den Workflow der Dokumentation integriert werden können und welchen Mehrwert sie gegenüber herkömmlichen Methoden besitzen und diese in Teilen obsolet machen.

Dies soll am Beispiel eines Surveys im kurdischen Teil des Irak aufgezeigt werden, der seit 2013 an der Universität Tübingen durch Peter Pfälzner durchgeführt wird. Ziel des Projektes ist die Identifikation und die Untersuchung von künstlichen Siedlungshügeln, sogenannten Tells, die meterhoch aus der Landschaft ragen und durch eine stetige Besiedlung desselben Ortes teils über mehrere Jahrtausende entstehen.

Besonders Beachtung sollen in diesem Beitrag die spezifischen Anforderungen der Archäologie des Vorderen Orients sowie die Auswirkungen der extremen äußeren Bedingungen auf die Arbeiten finden.

Dokumentation der Fundstellen

Eine der wichtigsten Aufgaben bei der Dokumentation der Tells liegt in einer genauen geographischen Verortung und einer möglichst präzisen Kartierung des gesamten Befundes. Diese Informationen können dann in einem nächsten Schritt in elektronischen Geoinformationssystemen verwaltet und analysiert werden, um beispielweise Rückschlüsse zur Siedlungsstruktur einer ganzen Region in einer spezifischen Epoche zu ermöglichen.

Aufgrund der großen Anzahl von Fundstellen im Untersuchungsareal muss diese Erfassung schnell und effizient erfolgen. Zusätzlich erschwert werden die Arbeiten durch die teilweise große Entfernung der einzelnen Sites, durch das häufig unwegsame Gelände und durch das vollständige Fehlen von bekannten Vermessungspunkten. Um diesen Problemen zu begegnen wurden verschiedene Ansätze evaluiert.

Herkömmliche Vermessungsmethoden

Eine tachymetrische Erfassung der Befunde ist aufgrund des Fehlens korrespondierender Festpunkte mit bekannter Koordinate nur in einem lokal beschränkten Vermessungsnetz möglich. Zudem ist bei dieser Methode aufgrund der Morphologie der Tells ein häufiges Umstationieren des Tachymeters

erforderlich, da stets eine Sichtverbindung zwischen Gerät und Winkelprisma notwendig ist. Darüber hinaus muss der Vermesser jeden Punkt der zu vermessenden Struktur anlaufen, um punktgenaue Daten zu erheben. Aufgrund der Geländestruktur war dies bei einem Großteil der zu vermessenden Sites nicht zu gewährleisten.

Somit muss diese Methode als ineffizient und aufgrund der fehlenden absoluten Verortung in einem standardisierten Koordinatensystem als ungeeignet angesehen werden.

Als zweite Methode wurde die Vermessung der Sites mittels GPS angedacht. Dies erlaubt zwar eine Vermessung in einem absoluten Koordinatensystem, nach wie vor muss jedoch jeder Punkt zeitaufwendig angelaufen werden, um eine präzise Kartierung zu gewährleisten.

Als weiterer Nachteil beider Methoden muss zudem angeführt werden, dass sie lediglich punktförmige Daten produzieren, die in einem zweiten Schritt erst aufwendig zu fertigen Karten zusammengeführt werden müssen.

Einsatz von luftgestützter Photogrammetrie

Aus diesen Gründen haben wir uns dazu entschieden, ausschließlich luftgestützte Photogrammetrie zur Vermessung der Siedlungshügel zu verwenden, die eine effiziente und präzise Erfassung erlaubt.

Setup und Einsatz der Drohne

Grundlage dieser Methoden sind Luftbilder sowie mittels GPS eingemessene Passpunkte. Zur Aufnahme der Bilder kam ein Quadrocopter des Typs *DJI Phantom Vision +* zum Einsatz, der sich per GPS positioniert und mit einer Funkfernsteuerung sowie einer Smartphone-App gesteuert wird. In dieser App sind ein Livebild der eingebauten Kamera sowie Telemetriedaten wie Flughöhe und Akkustand verfügbar. Aufgrund der hellen Umgebung waren diese Daten jedoch nur bedingt ablesbar und eine Positionierung der Drohne musste weitestgehend über einen Sichtkontakt zum Gerät selber erfolgen. Hier sollen in Zukunft verschiedene Schutzvorrichtungen vor zu starker Sonneneinstrahlung evaluiert werden. Die Aufnahme der Bilder erfolgt intervallgesteuert, so dass sich der Pilot lediglich um die Positionierung der Drohne über dem Grund zu kümmern hat. Als geeignete Flughöhe haben sich ca. 50 Meter erwiesen, die sowohl eine große Überlappung der Bilder als auch eine entsprechende Auflösung gewährleisten. Aufgrund der GPS-Positionierung der Drohne kann diese Flughöhe ohne Mühe konstant gehalten werden.

Um ein späteres Zusammenfügen der Bilder positionsgenau zu ermöglichen sind auf dem Grund Passpunkte verteilt, die mittels eines GPS-Gerätes genau eingemessen werden. Die Bilder der Drohne tragen zwar ebenfalls GPS-Informationen, diese zeigen jedoch die genaue Position der Drohne bei der Aufnahme an und nicht die Position des aufgenommenen Areals.

Als problematisch hat sich die Auswirkung großer Hitze auf die Elektronik der Drohne erwiesen, Abbrüche des Funkkontaktes zwischen Fernsteuerung und Drohne waren die Folge. Ein Schutzmechanismus lässt in diesem Fall den Kopter jedoch zu seinem Startpunkt zurückkehren, so dass hier ein sicherer Betrieb stets gewährleistet ist.

Erstellung von kartographischen Informationen aus Luftbildern mittels SFM

In einem zweiten Schritt müssen nun die Luftbilder prozessiert und mit den GPS-Koordinaten verbunden werden. Hierzu wird die Software Agisoft Photoscan Pro verwendet, die eine weitestgehend automatisierte Verarbeitung der Bildinformationen mittels Structure from Motion (SFM) zu einem fertigen 3D-Modell erlaubt. Hierzu werden in den Bildern gemeinsame Strukturen wie Eckpunkte oder Linien durch die Software erkannt und in einem dreidimensionalen Raum verortet. Hierzu ist eine Überlappung der verwendeten Bilder notwendig. Der Methode liegen dieselben

Prozesse zu Grunde, die das menschliche Gehirn zur Konstruktion dreidimensionaler Informationen verwendet.

In einem nächsten Schritt werden diese Passpunkte trianguliert und vernetzt. Dieses sogenannte Mesh kann nun wiederum mit einer photorealistischen Textur versehen werden, die ebenfalls aus den Luftbildern abgeleitet wird. Die auf den Bildern zu identifizierenden Passpunkte müssen mit den GPS-Koordinaten verbunden werden, um das Modell mit absoluten Größeninformationen zu verbinden. Ohne diese Zusatzinformationen ist das Modell maßstabslos und kann kaum sinnvoll genutzt werden.

Um nun aus diesem Modell eine Karte zu erzeugen, ist ein Export in ein Digital Elevation Modell (DEM) notwendig. Dieses besteht aus einem Rasterbild, in dem die Höheninformationen der einzelnen Pixel durch unterschiedliche Grautöne repräsentiert werden, die mit absoluten Höhenmetern verbunden sind. Die Ausdehnung sowie die Lage des DEMs sind durch die absolute Koordinaten in einem vorher zu bestimmenden Koordinatensystem definiert, die in einer Zusatzdatei abgelegt werden. Das DEM und diese Zusatzdatei können nun mit nahezu jedem Geoinformationssystem eingelesen und weiterverarbeitet werden. Zur Erstellung einer topographische Karte dient nun eine farbliche Repräsentation des DEM sowie hieraus abgeleitete Konturlinien.

Neben einem Export als DEM bedient Photoscan auch weitere Datenformate wie OBJ oder Collada, die eine Weiterverarbeitung des Modells in beliebigen 3D-Umgebungen gestattet.

Als Nachteil der hier präsentierten Methode muss das Fehlen jeglicher Features gelten, die sich im 3D-Modell nicht erkennen lassen. So sind Straßen, Wege sowie unterschiedliche Landnutzungen im DEM nahezu unsichtbar. Photoscan erlaubt jedoch neben dem Export als DEM auch gleichzeitig die Erstellung eines Orthofotos, das eine koordinatenrichtige, rechtwinklige Aufsicht auf das Modell darstellt. Dieses Orthofoto kann ebenfalls im Geoinformationssystem verarbeitet werden und die im Geländemodell unsichtbaren Informationen lassen sich dort nun digitalisieren und gemeinsam mit der topographischen Karte visualisieren.

Erfahrungen aus Ausblick

Die hier vorgestellte Methode hat sich bei der Dokumentation der Siedlungshügel überaus gut bewährt und muss als deutlich effizienter als die herkömmlichen Vermessungsmethoden gelten. Diese Aussage lässt sich jedoch nicht verallgemeinern. So ist eine derartige Vorgehensweise in einer dichten Vegetation deutlich schwieriger einzusetzen, als in der wüstenähnlichen Landschaft des Nordirak. Ebenso hat sich die dünne Besiedlung des Areal als großer Vorteil erwiesen, in einem dichtbesiedelten Gebiet ist der Einsatz solcher Systeme aufgrund von Aspekten der Sicherheit durchaus fragwürdig.

Als weiterer großer Vorteil der Methode kann die leichte Weiterverarbeitung der Daten in Geoinformationssysteme sowie 3D-Anwendungen angesehen werden, wodurch in einem einzelnen Arbeitsschritt vielfältige Anforderungen erfüllt werden können.

Die erreichbare Präzision hängt sehr stark von der Auflösung der Kamera, der Flughöhe sowie der Genauigkeit der Vermessung der Passpunkte ab. Für die großflächige Vermessung der Tells ist die erreichte Genauigkeit jedoch absolut ausreichend. In der kommenden Kampagne soll nach Möglichkeit einer zurzeit im Test befindliche, deutlich größere Drohne mit einem hochauflösenden Kamerasetup verwendet werden, das eine Bodenauflösung von 1-2 Zentimetern erlaubt und auch eine Kartierung kleinteiliger Befunde erlaubt.