Forschungsdaten in den

FORGE 2015

Geisteswissenschaften

Programm und Abstracts

gwin Geisteswissenschaftliche Infrastruktur für Nachhaltigkeit

http://uhh.de/forge2015

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Forschungsdaten in den Geisteswissenschaften (FORGE 2015)

vom 16. bis 18. September 2015 an der Universität Hamburg

Programmkomitee
Jonathan Blumtritt (DCH Köln)
Hanna Hedeland (HZSK, Hamburg)
Daniel Jettka (HZSK, Hamburg)
Maxi Kindling (HU Berlin)
Timm Lehmberg (HZSK, Hamburg)
Jan Christoph Meister (heureCLÉA, Hamburg)
Hagen Peukert (gwin, Hamburg)
Thomas Schmidt (IDS, Mannheim)
Stefan Schmunk (DARIAH-DE, Göttingen)
Stefan Thiemann (gwin, Hamburg)
Thorsten Trippel (CLARIN-D, Tübingen)
Iris Vogel (gwin, Hamburg)
Peter Wittenburg (RDA, EUDAT, Garching)
Stefan Winkler-Nees (DFG)
Kai Wörner (gwin, Hamburg)

Organisation
Hagen Peukert
Stefan Thiemann
Iris Vogel
Kai Wörner

Geisteswissenschaftliche Infrastruktur für Nachhaltigkeit
Universität Hamburg
Fakultät für Geisteswissenschaften
Edmund-Siemers-Allee 1
20146 Hamburg
gwin@uni-hamburg.de

Mittwoch, 16.09.2015

08:30-09:00		
09:00-09:30	**WORKSHOP:** Humanities Data Centre: Langzeitverfügbarkeit geisteswissenschaftlicher Forschungsdaten als organisatorische Herausforderung beim Aufbau und Betrieb eines Forschungsdatenzentrums.	
09:30-10:00		
10:00-10:30		WORKSHOP: Einführung in MAXQDA
10:30-11:00		
11:00-11:30		
11:30-12:00		
12:00-12:30		
12:30-13:00		
13:00-13:30	**Kaffeepause**	
13:30-14:00	**Begrüßung** Dr. Martin Hecht, Kanzler der Universität Hamburg Dr. Stefan Thiemann, CIO der Fakultät für Geisteswissenschaften	
14:00-14:30	**Session 1. Chair: Stefan Thiemann** Forschungsdaten in historischen Projekten oder die Quelle im Netz [Daniel Burckhardt, Anna Menny, **S. 31**]	
14:30-15:00	Database of historical places, persons, keywords, and lemmas [Natalia Korchagina, **S. 19**]	
15:00-15:30	histHub.net - Ein mehrsprachiger Metadatenhub für Linked Open Data zu historischen Entitäten aller Epochen und Regionen der Schweiz [Pascale Sutter, **S. 42**]	
15:30-16:00	**Kaffeepause**	
16:00-16:30	**Session 2. Chair: Kai Wörner** Prämissen und Schlussfolgerungen der Angebotsentwicklung des Humanities Data Centres mit Hilfe des Konzepts des „Forschungsdatentyps“ [Andreas Aschenbrenner, Stefan Buddenbohm, Sven Bingert, Claudia Engelhardt, Ulrike Wuttke, **S. 3**]	
16:30-17:00	Vom analogen Archiv zum digitalen Forschungsdatenzentrum [Thomas Schmidt, **S. 40**]	
17:00-17:30	Vom digitalen Hausmeister zum international vernetzten Forschungsdatenzentrum. Die Entwicklung der digitalen Infrastruktur des Hamburger Zentrums für Sprachkorpora [Daniel Jettka, Hanna Hedeland, Timm Lehmberg, **S. 15**]	

Donnerstag, 17.09.2015

08:30-09:00	**Kaffepause**
09:00-09:30	**Postersession, Systempräsentationen und Fachmesse**
09:30-10:00	
10:00-10:30	Kaffee gesponsert von der programmfabrik
10:30-11:00	
11:00-11:30	
11:30-12:00	
12:00-12:30	**Mittagspause**
12:30-13:00	
13:00-13:30	**Kaffeepause.**
13:30-14:00	**Session 3. Chair: Timm Lehmberg** IANUS als fachspezifisches Forschungsdatenzentrum für die Altertumswissenschaften in Deutschland [Felix Schäfer, Maurice Heinrich, **S. 35**]
14:00-14:30	Überlegungen zu einem Betriebsmodell für etablierte wissenschaftliche Software [Thomas Schmidt, Kai Wörner, **S. 37**]
14:30-15:00	Die Entwicklung eines Geschäftsmodells für den Betrieb der Virtuellen Forschungsumgebung FuD. Ein Erfahrungs- und Praxisbericht [Marina Lemaire, Gisela Minn, Yvonne Rommelfanger, Thomas Burch, **S. 28**]
15:00-15:30	**Kaffeepause**
15:30-16:00	**Session 4. Chair: Hagen Peukert** Knora: A Generic Storage System and API for Accessing and Reusing Digital Qualitative Research Data in the Humanities [Benjamin Geer, Tobias Schweizer, Lukas Rosenthaler, Ivan Subotic, **S. 13**]
16:00-16:30	Bereitstellung und Nachnutzung von Forschungsdaten in den Geisteswissenschaften [Benedikt Fecher, **S. 7**]
16:30-17:00	Objekterfassung, Präsentationsportal, Forschungsumgebung – die Bereitstellung von musealem Content für Forschung und Wissenschaft [Dr. Michael Merkel, Frank Dührkohp, **S. 6**]
17:00-17:30	Kuration: Migration und Qualitätssicherung digitaler Forschungsdaten [Iris Vogel, **S. 44**]

Freitag, 18.09.2015

08:30-09:00	**Kaffeepause**
09:00-09:30	**Session 5. Chair: Iris Vogel** GeSIG – ein Projekt im Spannungsfeld einer sich entwickelnden Infrastrukturlandschaft [Cordula Meißner, Franziska Wallner, **S. 46**]
09:30-10:00	Charakteristika sprachtypologischer Daten und ihre projektübergreifende Integration mit Hilfe von digitalen Ontologien [Alexander Borkowski, Andrea C. Schalley, **S. 1**]
10:00-10:30	Relationale Datenbank und Webtechnologie im Dienst der Geolinguistik: eine Projektvorstellung [Susanne Oberholzer, Markus Kunzmann, **S. 33**]
10:30-11:00	INEL – ein Infrastrukturvorhaben zur Dokumentation indigener nordeurasischer Sprachen [Beáta Wagner-Nagy, Hanna Hedeland, Timm Lehmberg, Michael Rießler, **S. 21**]
11:00-11:30	**Kaffeepause**
11:30-12:00	**Session 6. Chair: Kai Wörner** Empirische Daten für die Kunstgeschichte erzeugen mit Artigo Social Image Tagging [Harald Klinke, **S. 18**]
12:00-12:30	CLLD - Dateninfrastruktur für Diversity Linguistics [Robert Forkel, **S. 10**]
12:30-13:00	Wissenstransfer und Wissensressourcen: Support und Helpdesk in den Digital Humanities [Timm Lehmberg, **S. 25**]
13:00-13:30	**Verabschiedung**

Vorträge

Charakteristika sprachtypologischer Daten und ihre projektübergreifende Integration mit Hilfe von digitalen Ontologien

Alexander Borkowski [a.borkowski@griffith.edu.au] Griffith University, Australien
Andrea C. Schalley [a.schalley@griffith.edu.au] Griffith University, Australien

Zusammenfassung

Sprachtypologische Forscher arbeiten mit Ton- und Filmmaterial angereichert durch graphische und textuelle Informationen. Letztere werden in einem Spektrum von unstrukturierten bis hin zu semistrukturierten Formaten erstellt und unterliegen einer weitgehend impliziten fachspezifischen Interpretation. Diese Mischung gepaart mit forscher- und projektspezifischen Schwerpunkten stellt eine Herausforderung dar für den nachhaltigen Umgang mit sprachtypologischen Daten und die projektübergreifende Integration dieser.

Im Rahmen des vom Australian Research Council (ARC) geförderten Projekts „Social Cognition and Language – The design resources of grammatical diversity" sowie kleinerer Folgeprojekte wurde ein Verfahren entwickelt, welches auf digitalen Ontologien basiert und eben jene Integration ermöglichen soll. Wir wollen in unserem Beitrag näher auf die Charakteristika sprachtypologischer Daten – insbesondere die von semistrukturierten Annotationen – in Bezug auf Ontologien eingehen und die resultierende Architektur einer Software für sprachtypologische Datenspeicherung und Weiterverarbeitung vorstellen.

Abstract

Die (Sprach-)Typologie als Unterdisziplin der Sprachwissenschaften hat das Ziel, Sprachen anhand ihrer formalen Charakteristika zu vergleichen und zu klassifizieren. Die Arbeitsweise von Typologen umfasst, grob umrissen, a) die Konzeptionierung von sprachunabhängigen Methoden zur Erhebung von sprachlichen Äußerungen (oft basierend auf visuellen Stimuli, etwa Bilder oder Videos), b) die Audio- und vermehrt Videoaufnahme von sprachlichen Äußerungen im Feld, c) die Annotation der Aufnahmen unter Verwendung von spezialisierter Software (ELAN, EXMARaLDA; cf. Sloetjes & Wittenburg, 2008; Schmidt & Wörner, 2014), d) den Vergleich der erhobenen Daten mit denen anderer Sprachen, etwa aus der Literatur, und e) die Veröffentlichung der Ergebnisse.

Ein großer Anteil der typologischen Forschungsergebnisse findet sich in gedruckter Form in einzelsprachlichen Darstellungen. Zudem gibt es wenige gedruckte Grammatiken, die durch ihre Strukturierung einen direkten Sprachvergleich unterstützen (wie dies die *Routledge Descriptive Grammar Series* auf Basis von Comrie & Smith's (1977) *Lingua Descriptive Studies Questionnaire* anstrebt). Diese sind nach Abschluss eines jeweiligen Buchprojektes jedoch weder an spätere Ergebnisse anpassbar noch grundsätzlich erweiterbar. Es gib einige wenige digitale Systeme, die versuchen, Ergebnisse aus möglichst vielen Sprachen zusammenzuführen (WALS, TDS, cf. Dryer & Haspelmath, 2013; Dimitriadis et al., 2009). Allerdings können diese dann aber generell keine Antworten auf spezifische, bislang ggfs. noch nie gestellte, Fragen liefern.

Die Problemstellung der Typologie leitet sich aus mehreren Gegebenheiten her. So wird die Anzahl der im Moment noch lebenden Sprachen auf ca. 7000 geschätzt, welche stark diversifiziert sind. Forscher haben also grundsätzlich die Wahl, sich auf eine begrenzte Anzahl von Sprachen, eine begrenzte Anzahl von zu untersuchenden Phänomenen und/oder eine begrenzte Auswahl an Literatur bzw. Feldforschungsdaten zu konzentrieren. Es ist schlichtweg unmöglich, dass Einzelne sich mit der notwendigen Vielzahl von Sprachen tiefgehend genug beschäftigen, um globale Vergleiche anstreben zu können.

Ferner sind die erhobenen Primärdaten (also Ton- und Bilddaten) nicht automatisiert typologisch zu analysieren (anders als etwa Sensordaten als Eingabe für metereologische Modellrechnungen) – es bedarf der Annotation durch Experten. Diese Annotationen sind komplexe, vielschichtige und zugleich lückenhafte (bedingt etwa durch mangelndes Wissen über die Gegenstandssprache oder fehlende Relevanz einzelner Aspekte für den aktuellen Forschungsschwerpunkt) Konstrukte in semistrukturierter Form, welche einer impliziten fachspezifischen Interpretation unterliegen. Erschwerend sind die den Sprachbeschreibungen zugrunde liegenden Terminologien aufgrund der Vielfalt an Sprachen oft nicht standardisiert sondern individuell geprägt, was auch digitalen Suchverfahren einen repositorienübergreifenden Zugang erschwert. Für Maßnahmen zur Sicherstellung von Einheitlichkeit (etwa doppelte Annotation mit Abgleich) fehlen Zeit, Geld sowie Anreize. Genau hier setzt der von uns vorgestellte Lösungsansatz an. Primär- und Sekundärdaten werden in einem fachspezifischen Archiv vorgehalten, welches Mehrwertdienste und Unterstützung im Bezug auf Datenaufnahme, -verarbeitung und -integration bereitstellt und somit einen Anreiz bietet, bei der Annotation – also der Erstellung der Sekundärdaten – einen Mehraufwand zu betreiben. Ergebnis ist ein flexibler Datenspeicher, welcher Daten zu speziellen sprachtypologischen Fragestellungen vergleichend zugänglich machen kann und somit Zugriff auf die existierenden Informationen ermöglicht, seien sie von Spezialisten für Einzelsprachen in der Tiefe oder von Spezialisten für sprachübergreifende Phänome in der Breite bereitgestellt. Insbesondere sollen Suchanfragen nicht auf die objektsprachliche Formebene alleine beschränkt sein sondern können ebenso funktional/semantisch basiert sein, um etwa unterschiedliche einzelsprachliche Kodierungen von gleichen funktionalen Konzepten liefern zu können.

Ein Ziel des vom Australian Research Council (ARC) geförderten Projekts „Social Cognition and Language – The design resources of grammatical diversity“ war es, ein sprachübergreifendes digitales Modell davon zu erstellen, wie sich soziale Kognition in den Grammatiken von einzelnen Sprachen wiederfindet. Ein solches Modell stellt eine Verknüpfung von Konzepten dar, welche man auch als (konzeptuelle) Ontologie bezeichnen kann. Als technische Grundlage für ein geeignetes Sofwaresystem boten sich die entsprechenden digitalen Ontologien des Semantic Web an. Im Rahmen des Projekts wurde die Entwicklung der darauf basierenden Software TYTO („typology tool“) vorangetrieben, welche eine mögliche Umsetzung des vorgestellten Lösungsansatzes ist. In unserem Vortrag werden wir diesen Lösungsansatz darlegen sowie aufzeigen, wie er konkret die oben angeführten Desiderata und Probleme adressiert.

Comrie, B., & Smith, N. (1977). Lingua descriptive studies: Questionnaire. *Lingua* 42: 1–72.

Dimitriadis, A., Windhouwer, M., Saulwick, A., Goedemans, R., & Bíró, T. (2009). How to integrate databases without starting a typology war: the typological database system. In Everaert, M., Musgrave, S., & Dimitriadis, A. (eds.), *The Use of Databases in Cross-Linguistic Studies,* 155–207. Berlin: Mouton de Gruyter.

Dryer, M. S., & Haspelmath, M. (eds.). (2013). *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Schmidt, T., & Wörner, K. (2014). EXMARaLDA. In J. Durand, U. Gut, & G. Kristoffersen (eds.), *Handbook on Corpus Phonology,* 402–419. Oxford: Oxford University Press.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).* Marrakech: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Prämissen und Schlussfolgerungen der Angebotsentwicklung des Humanities Data Centres mit Hilfe des Konzepts des „Forschungsdatentyps“

Stefan Buddenbohm [buddenbohm@mmg.mpg.de], Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften
Claudia Engelhardt [claudia.engelhardt@sub.uni-goettingen.de], Niedersächsische Staats- und Universitätsbibliothek Göttingen
Ulrike Wuttke [uwuttke@gwdg.de], Akademie der Wissenschaften zu Göttingen

Zusammenfassung

Der Vortrag stellt die Angebotsdefinition für ein geisteswissenschaftliches Forschungsdatenzentrum am Beispiel des Humanities Data Centre (HDC) vor und geht dabei insbesondere auf das Konzept des Forschungsdatentyps näher ein.

Abstract

Das Projekt Humanities Data Centre[1][2] erarbeitet in der vom Niedersächsischen Ministerium für Wissenschaft und Kultur / Niedersächsischen Vorab geförderten Designphase (2014-2016) die konzeptionellen Grundlagen für den Aufbau und Betrieb eines Forschungsdatenzentrums für die Geisteswissenschaften. Die Designphase, in der die organisatorischen und technischen Grundlagen konzipiert und prototypisch umgesetzt werden, ist die Basis für den anschließenden Aufbau des Datenzentrums.

Geisteswissenschaftliche Forschungsdaten sollen über die üblichen Lebenszyklen von Projekten, Datenformaten und Software hinaus nutzbar sowie über die Anwesenheit der ursprünglichen Datenersteller hinaus interpretierbar und nachnutzbar sein. Diese Anforderung stellt sich sowohl aus der Sicht der Wissenschaftlerinnen und Wissenschaftler aus Gründen der Guten Wissenschaftlichen Praxis, der Nachnutzbarkeit von Forschungsdaten und als Teil des wissenschaftlichen Diskurses, als auch aus der Perspektive von Forschungsförderern mit Blick auf die Effizienz und Nachhaltigkeit von Projektfinanzierungen.

Bestimmte Charakteristika (geisteswissenschaftlicher) Forschungsdaten stehen aber einer Standardisierung, wie sie beispielsweise bei Publikationen gut möglich ist, entgegen und erhöhen damit entscheidend den Aufwand für ihre Langzeitarchivierung und langfristige Bereitstellung zur Nachnutzung:

- Forschungsdaten bestehen zunehmend aus verbundenen Objektsammlungen, etwa Datenbanken oder Applikationen (wie zum Beispiel Digitale Editionen und Visualisierungen), in deren Design bereits wissenschaftliche Erkenntnis steckt, gewinnen immer größere Bedeutung.

[1] www.humanities-data-centre.org
[2] Das Humanities Data Centre ist über diesen Vortrag hinaus mit einem Workshop im Vorfeld der Konferenz vertreten. In diesem Workshop werden v.a. organisatorische und Herausforderungen beim Aufbau eines geisteswissenschaftlichen Forschungsdatenzentrums diskutiert.

- Forschungsdaten und die Anforderungen an das Forschungsdatenmanagement sind in großen Teilen disziplinspezifisch. Oft ist nur ein administrativer Kern (von Metadaten) einfach standardisierbar, nicht aber die Inhalte an sich.
- Viele Dienste zur Archivierung und Bereitstellung von heterogenen Forschungsdaten skalieren nicht. Anders als bei stark standardisierten Daten ist die Archivierung und Nachnutzung von heterogenen Forschungsdaten stark vom Einzelfall abhängig.

Diese Charakteristika stark heterogener geisteswissenschaftlicher Forschungsdaten verlangen tendenziell disziplin- und fallspezifische Lösungen. Im Gegensatz dazu haben Infrastrukturanbieter, wie Rechenzentren und Bibliotheken, ein großes Interesse an der Standardisierbarkeit und Skalierbarkeit von Services, um die langfristige Finanzierbarkeit ihrer Angebote sichern zu können. So ist der Aufbau eines Forschungsdatenzentrums, das beständig lernt und seine Angebote (auch unter Berücksichtigung von Einzelfällen) erweitert, aus Sicht der Infrastruktureinrichtungen zwar technisch und organisatorisch denkbar, aber finanziell unrealistisch. Der Widerspruch zwischen der Notwendigkeit der kontinuierlichen Weiterentwicklung lokaler Angebote bei gleichzeitig stark eingeschränkter Skalierbarkeit der Angebote ist eine der großen Herausforderungen beim Aufbau eines Forschungsdatenzentrums.

Wie aber kann der Zielkonflikt zwischen der Heterogenität geisteswissenschaftlicher Forschungsdaten einerseits und Standardisierungs- und Skalierbarkeitsanforderungen innerhalb eines Forschungsdatenzentrums andererseits erfolgversprechend adressiert werden? Bei der Beantwortung dieser Fragen für das Humanities Data Centre standen folgende Punkte im Vordergrund:

- Was sind eigentlich geisteswissenschaftliche Forschungsdaten? Welche Formate, Objekte und Anwendungen kommen grundsätzlich für ein Forschungsdatenzentrum in Betracht?
- Welche Angebote erwarten Fachwissenschaftlerinnen und Fachwissenschaftler von einem Forschungsdatenzentrum?
- Was sind sinnvolle Kooperationsmodelle – sowohl zwischen den Fachwissenschaftlerinnen und Fachwissenschaftlern und dem Datenzentrum, als auch zwischen verschiedenen regionalen und überregionalen Infrastrukturanbietern?

Als Antwort auf diese Fragen und als Instrument zur Angebotserstellung hat das HDC das Konzept des Forschungdatentyps entwickelt. Als organisatorische (und technische) Alternative zur völlig individualisierten Behandlung von Forschungsprojekten entwickelt das HDC spezialisierte Angebote für verschiedene Cluster von Forschungsdatentypen. Dieses Vorgehen beruht auf der Erfahrung, dass wiederkehrende wissenschaftliche Methoden und Projektkontexte zu wiederkehrenden Mustern in den Forschungsdaten führen.

Das Konzept des Forschungsdatentyps geht von folgenden Grundannahmen aus:

- Die Angebote eines geisteswissenschaftlichen Forschungsdatenzentrums sind forschungsdatentypenspezifisch, das heißt für jeden Forschungsdatentyp müssen individuell unterschiedliche Dienste und Abläufe entwickelt werden. Geisteswissenschaftliche Forschungsdatenzentren können sich durch die Spezialisierung auf unterschiedliche Forschungsdatentypen unterscheiden.
- Daraus folgt, dass es nicht ein geisteswissenschaftliches Forschungsdatenzentrum geben sollte, sondern eine Reihe von Forschungsdatenzentren mit unterschiedlichen Spezialisierungen über ein generisches Grundangebot hinaus. Auf einem bestimmten Forschungsdatentyp basierende Angebote müssen nicht über die Forschungdatenzentren übergreifend standardisiert werden, allerdings wird vermutlich Konvergenz auf Grund von Angebot und Nachfrage entstehen. Mit anderen

Worten: Es wird nicht „ein“ oder „das perfekte“ Angebot für einen Forschungsdatentyp geben, sondern die unterschiedlichen geisteswissenschaftlichen Datenzentren werden auch individuelle Angebote haben.

- Obwohl sich die forschungsdatentypischen Angebote gemeinsam mit den Forschungsmethoden entwickeln müssen, ist davon auszugehen, dass sie sich weniger schnell entwickeln werden als zum Beispiel spezielle Datenformate oder Software. Sie bieten daher einen hinreichend stabilen Container für das Management von Migrationszyklen.

Im Vortrag wird die auf der Grundlage der oben genannten Prämissen erarbeitete Angebotsstruktur des Humanities Data Centres skizziert. Wir hoffen auf eine kritische Diskussion dieses Vorgehens und unserer Ergebnisse mit den zukünftigen Nutzerinnen und Nutzern, den Fachwissenschaftlerinnen und Fachwissenschaftlern, sowie Vertretern anderer Informationsinfrastruktureinrichtungen wie Rechenzentren und Bibliotheken.

Objekterfassung, Präsentationsportal, Forschungsumgebung

Die Bereitstellung von musealem Content für Forschung und Wissenschaft

Dr. Michael Merkel [Michael.Merkel@amh.de], Archäologisches Museum Hamburg
Frank Dührkohp [frank.duehrkohp@gbv.de], Verbundzentrale des GBV

Zusammenfassung

Der Vortrag gibt einen Überblick über den Stand der musealen Objekterfassung, die Vernetzung mit anderen Datenquellen über Standardschnittstellen sowie die Bereitstellung der Daten über standardisierte Austauchformate.

Abstract

In den Museen sind in den vergangen Jahren große Anstrengungen, unterstützt von Förderprogrammen der Forschungsförderer, Stiftungen aber auch der jeweiligen Träger, unternommen worden, die Sammlungen der Häuser zu inventarisieren, zu die Objekte digitalisieren und die Inhalte für regionale (Museen Nord), nationale (DDB) und internationale Portale (Europenana) zur Verfügung zu stellen. So konnten unterschiedliche Softwarelösungen, standardisierte Datenformate und umfangreiche Digitalisierungsworkflows entwickelt und zur Verfügung gestellt werden. Allerdings lag der Focus der Entwicklung im Rahmen EDV-Erfassung von Museen vor allem auf der Inventarisierung des eigenen Bestandes sowie auf der technischen Unterstützung der musealen Verwaltungsworkflows (Ein- und Ausgang, Leihverkehr, Restaurierung), nicht aber auf der Bereitstellung der Daten für Forschungsumgebungen oder der Vernetzung mit externen Datenquellen. Im Zuge aktueller Entwicklungen und vor allem in Bezug auf den Bedarf der Bereitstellung der musealen Metadaten für Forschungsrepositorien, stehen die Museen vor neuen Herausforderungen:

1. Die Bereitstellung von Wissensquellen anderen Einrichtungen und der relevanten Nachbardisziplinen erfordert eine Vernetzung der Systeme über standardisierte Schnittstellen.
2. Der Trend zu Spezialdatenbanken innerhalb der musealen Erfassung erfordert eine Interoperabilität dieser Datenbanken in Bezug auf die datenbankübergreifende Suche.
3. Die Nachnutzung der Daten in Forschungsinfrastrukturen erfordert eine Übergabe der Metadaten in standardisierten Austauschformaten sowie die Bereitstellung über Harvesting-Schnittstellen.
4. Die wissenschaftliche Inventarisierung von Spezialsammlungen durch spezialisierte Wissenschaftler erfordert eine Öffnung der Museumsdatenbanken durch Webinterfaces sowie die Möglichkeit einer kollaborativen Erfassung eines gemeinsamen Datenbestandes durch Kuratoren unterschiedlicher Museumssammlungen.

Der Vortrag soll einen Überblick über den aktuellen Stand dieser Entwicklungen geben sowie diese protoptypisch am Beispiel eines neuen Dienstes des Archäologischen Museums Hamburg in Kooperation mit der Verbundzentrale des GBV vorstellen.

Bereitstellung und Nachnutzung von Forschungsdaten in den Geisteswissenschaften

Potenziale der Sekundärdatennutzung in den Geistes- und Kulturwissenschaften

Benedikt Fecher [fecher@hiig], Alexander von Humboldt Institut für Internet und Gesellschaft, Berlin und Deutsches Institut für Wirtschaftsforschung, Berlin

Zusammenfassung

Dieser Beitrag behandelt die Bereitstellung und Nachnutzung von Forschungsdaten. Hierzu werden Ergebnisse einer Befragung unter Forschenden vorgestellt und Potenziale und Barrieren für die Kultur- und Geisteswissenschaften erörtert.

Abstract

Die Bereitstellung und Nachnutzung von Forschungsdaten birgt großes Potenzial für den wissenschaftlichen Fortschritt. Mit "alten" Daten lassen sich, ohne eigenen Erhebungs- und Dokumentationsaufwand, neue Forschungsfragen beantworten. Zudem ermöglicht der offene Zugang zu Forschungsdaten Replikationsstudien und somit die intersubjektive Nachvollziehbarkeit von Ergebnissen. Offene Forschungsdaten tragen somit auch zur Qualitätssicherung von Forschung bei. Darüber hinaus ermöglichen sie die Erprobung neuer Erhebungs- und Analysemethoden im, wie beispielsweise Data- und Text-Mining oder semantische Textanalysen. Neelie Kroes, die frühere EU-Kommissarin für die Digitale Agenda, geht sogar davon aus, dass in der Bereitstellung und Nachnutzung von Forschungsdaten das Potenzial steckt, Europas Innovationsleistung zu fördern (1).

Es ist daher kaum verwunderlich, dass Forschungsförderer und wissenschaftliche Organisationen wie die Deutsche Forschungsgemeinschaft oder die Allianz der deutschen Wissenschaftsorganisationen die Bereitstellung von Forschungsdaten zunehmend fordern (2, 3).

Selbst Forschende erkennen das große Potenzial von Sekundärdaten für die eigene Forschung und deren Bedeutung für den wissenschaftlichen Fortschritt; sie stellen aber nur äußerst selten selbst Forschungsdaten bereit (4). Eine besonders geringe Kultur der Bereitstellung und Nachnutzung lässt sich in den Geistes- und Kulturwissenschaften feststellen (5).

Es stellt sich die Frage, wo die Potenziale offener Daten in den Geistes- und Kulturwissenschaften liegen und was Forschende davon abhält, ihre Daten bereitzustellen.

Angesichts des großen Potenzials offener Forschungsdaten für den wissenschaftlichen Fortschritt stellt sich die Frage, was Forschende, und speziell Forschende in den Geistes- und Kulturwissenschaften, von der Bereitstellung und Nachnutzung von Forschungsdaten abhält. Weiterhin stellt sich die Frage, inwiefern die Kultur- und Geisteswissenschaften von offenen Forschungsdaten und digitaler Dateninfrastruktur profitieren können.

In diesem Vortrag werden Ergebnisse einer Befragung von n = 1542 akademisch Forschenden in Deutschland präsentiert. Die Befragung wurde im Rahmen des Leibniz Forschungsverbunds Science 2.0 im Herbst 2014 durchgeführt. Der Vortrag beinhaltet aktuelle Zahlen zum Umgang mit und der Nachnutzung von Forschungsdaten sowie Barrieren und Anreize, Daten bereitzustellen. Ein besonderes Augenmerk liegt auf den geistes- und kulturwissenschaftlichen Fächern. Mit SowiDataNet wird zudem ein Repositorium-Projekt der Leibniz-Gemeinschaft vorgestellt. In der anschließenden Diskussion können Potenziale und Herausforderungen des Forschungsdatenaustausches in den Geistes- und Kulturwissenschaften erörtert werden.

1. N. Kroes, Opening Science Through e-Infrastructures (2012), (available at http://europa.eu/rapid/press-release_SPEECH-12-258_en.htm).
2. W. L. I. (LIS) DFG, "Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten" (Richtlinie, Bonn, 2009), (available at http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf).
3. Allianzinitiative, Grundsätze zum Umgang mit Forschungsdaten, (available at http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze.html).
4. C. Tenopir *et al.*, Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE.* **6**, e21101 (2011).
5. B. Fecher, S. Friesike, M. Hebing, S. Linek, A. Sauermann, A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing. *DIW*

Berlin Discussion Paper. **1454** (2015) (available at http://dx.doi.org/10.2139/ssrn.2568693).

Cross-Linguistic Linked Data

Dateninfrastruktur für Diversity Linguistics

Robert Forkel [forkel@shh.mpg.de] Max-Planck-Institut für Menschheitsgeschichte

Zusammenfassung

Dieser Vortrag stellt das Cross-Linguistic Linked Data Projekt (CLLD) [1] vor und beschreibt zwei originelle Praktiken, die in diesem Projekt entwickelt wurden, um Forschungsdaten kollaborativ zu pflegen und die nachhaltige Nutzbarkeit von Daten und Anwendungen zu ermöglichen.

Abstract

Das Forschungsgebiet, das hier mit Diversity Linguistics [2] bezeichnet wird, umfasst Deskriptive Linguistik, Typologie und Historische Linguistik und ist generell durch die Beschäftigung mit (möglichst) vielen der etwa 7500 Sprachen der Welt [3] gekennzeichnet.

Diese Charakteristik hat unmittelbare Auswirkung auf die Art der Forschungsdaten, die eine Rolle spielen: Sobald die Betrachtung von mehreren hundert Sprachen im Vordergrund steht, reduziert sich das verfügbare Material meist auf Wortlisten, Phoneminventare, typologische Surveys oder kleine Sammlungen von Beispielsätzen. Als Beispiele für große Datensammlungen in diesem Feld können WALS (The World Atlas of Language Structures) [4] oder ABVD (Austronesian Basic Vocabulary Database) [5] gelten.

Die Herausforderungen an Forschungsdateninfrastrukturen liegen hier also weniger darin, mit grossem Datenvolumen zurechtzukommen (ABVD umfasst momentan 237.921 lexikalische Einträge, WALS enthält 76.465 Datenpunkte), als vielmehr darin, möglichst viele der Daten die bereits gesammelt wurden, möglichst sinnvoll zugänglich zu machen, sowie best practices für Datenpflege und -sammlung zu entwickeln.

Dieser Aufgabe stellt sich das CLLD Projekt. Durch das Bereitstellen einer Publikationsplattform sollen Datenpublikationen einerseits technisch vereinfacht werden, andererseits ein größeres Prestige bekommen, so dass weder Angst vor technischen Schwierigkeiten noch die Befürchtung, um die Früchte der eigenen Arbeit gebracht zu werden, als Argument gegen eine Veröffentlichung der Daten angeführt werden können.

Um die Chancen einer digitalen Publikation im Web voll zu nutzen, soll es natürlich möglich sein, Datenbanken weiter zu pflegen, zu ergänzen und zu verbessern. Mit diesem Anspruch einher geht das Problem kollaborativer Datenpflege und der nachhaltigen Verfügbarkeit mehrerer Bearbeitungsstände von Datensätzen. Zwei interessante Lösungen für diese Probleme, die die Besonderheit der Daten im betrachteten Forschungsfeld berücksichtigen, werden im Folgenden vorgestellt.

Das source-code-management tool git [6] – insbesondere in Verbindung mit der hosting Plattform GitHub [7] – zur kollaborativen Pflege von Forschungsdaten zu verwenden, ist sicher keine ganz neue Idee [8]. Für die Art Daten, die oben beschrieben wurde, ist einworkflow, der auf GitHub basiert aber geradezu ideal, weil

- die Datenmenge meist problemlos in einem repository untergebracht werden kann
- und bei geigneter Formatwahl (etwa CSV), Zeilenbasierte diffs von Text-Datensehr hilfreich sind.

Versionsverwaltung und diffs sind aber nicht die einzigen Werkzeuge aus dem Bereich der Softwareentwicklung, die "natürliche" Entsprechungen im Forschungsdaten-management haben:

- "pull requests" bieten einen Mechanismus für begutachtete Daten-updates.(Siehe https://github.com/clld/tsammalex-data/pull/8)
- Ein "fork" ermöglicht einen transparenten Transfer von Verantwortlichkeiten.
- Ein "release" stellt die formale Publikation einer Datenbank dar.

Aber GitHub und best practices in der Softwareentwicklung bieten noch mehr:

- Die Anbindung an "continuous integration" services wie Travis-CI macht esmöglich, nach jeder Änderung der Daten Konsistenzprüfungen durchführen zu lassen. (Siehe https://travis-ci.org/clld/tsammalex-data/builds)
- Die Integration mit zenodo [9] erlaubt automatisches Archivieren von releases.

Die Integration von GitHub mit zenodo kann aber nicht nur zur (zitierfähigen)Archivierung von Datenpublikationen genutzt werden, sondern kann auch eine zentrale Rolle spielen, wenn es darum geht, Applikationen nachhaltig verfügbar zu machen [10].

Viele der Datenbanken, die auf der CLLD Plattform publiziert werden (unter anderem WALS), werden durch eine massgeschneiderte Web-Applikation zugänglich gemacht undverfolgen ein editions-basiertes Publikationsmodell. Das Problem, ältere Editionen einer Datenbank durch die zugehörige Applikation zugänglich zu machen, kann folgendermassen gelöst werden:

1. Aktuelle releases der Datenbank werden im source code repository derApplikation gepflegt.
2. Die Applikation verfügt über Funktionalität, diese Daten zu importieren.
3. Im source code repository ist ein bootstrap script verfügbar, mit dem Installation, Import und Start der Applikation automatisiert werden.

Zugriff auf ältere Editionen einer CLLD Datenbank sieht dann wie folgt aus:

1. Der von zenodo für den release des GitHub repositories vergebene DOI (etwa 10.5281/zenodo.11040 für release v2014.2 von clld/wals3) führt zum archivierten repository.
2. Nach download und Entpacken dieses Archivs wird das bootstrap scriptausgeführt, das die Applikation lokal startet.

Im Idealfall von "reproducible research" erfolgte der Zugriff auf die Daten für eine Analyse automatisiert und über eine öffentliche API. Dann kann diese Analyse - nach Anpassen des API-providers auf die lokale Umgebung - mit der "wieder-belebten" Datenbank reproduziert werden.

Um diesen Mechanismus mittel- oder langfristig nutzbar zu machen, ist allerdings noch weitere Infrastrukur erforderlich, etwa Virtualisierungslösungen, die die nötige System-umgebung zur Verfügung stellen. Institutionell gepflegte Docker repositories [11] oder

EC2 AMI [12] Sammlungen wären hier ein erster Schritt. Die Aufgabe solche Infrastrukur zu schaffen und bereitzustellen, scheint aber ohne weiteres im Portfolio eines Rechenzentrums oder einer Digitalen Bibliothek vorstellbar. Ebenso vorstellbar ist es, dass Digitale Bibliotheken auch den nächsten Schritt gehen, und on-demand Datenbanken zu Verfügung stellen, obigem Szenario folgend etwa als Service der für geeignete DOIs abgerufen werden kann.

Zusammenfassend kann man feststellen: Wenn es die Art der Daten erlaubt, steht mit GitHub heute eine sehr gute Plattform zuf Verfügung, um best practices von der Softwareentwicklung ins Datenmanagement zu übertragen. Das Problem der Abhängigkeit von GitHub als Service wird dadurch abgemildert, dass alle Daten, die in diesen workflows gepflegt werden, sowie ein Großteil der Metadaten (commit messages, etc.) in jedem einzelnen repository clone vorhanden sind, und damit einfach in lokale backup-wprkflows integriert werden können.

[1] http://clld.org

[2] http://dlc.hypotheses.org/about

[3] http://glottolog.org

[4] http://wals.info

[5] http://language.psy.auckland.ac.nz/austronesian/

[6] http://git-scm.com

[7] https://github.com

[8] http://clld.org/2015/02/03/open-source-research-data.html

[9] https://zenodo.org

[10] http://clld.org/2014/07/28/citing-clld-databases.html

[11] https://hub.docker.com/

[12] http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AMIs.html

Knora

A Generic Storage System and API for Accessing and Reusing Digital Qualitative Research Data in the Humanities

Benjamin Geer [benjaminlewis.geer@unibas.ch], Digital Humanities Lab, University of Basel

Tobias Schweizer, [t.schweizer@unibas.ch], Digital Humanities Lab, University of Basel

Summary

While an increasing volume of primary source materials and research data in the humanities is now available in online repositories, it is often difficult to work with this data in new research. Facilities for searching, annotating, linking, and extending data produced by different projects, and stored in different repositories, are limited or nonexistent, or at best highly heterogeneous. Moreover, the great diversity of storage systems underlying existing repositories makes it cumbersome to construct such facilities at present. We argue that these problems can be solved by storing research data in RDF triplestores, using a flexible set of shared ontologies that abstract out the basic commonalities in diverse data sets, along with project-specific ontologies derived from these, and by providing a generic, extensible Application Programming Interface (API) for accessing the data. We describe Knora (Knowledge Organization, Representation, and Annotation), an implementation of this approach, and discuss the lessons learned from its development and use.

Abstract

Despite the proliferation of online repositories of primary source materials and research data in the humanities, it is still difficult for researchers to use this data when doing their research. Many repositories are simply collections of web pages that provide images or PDF files for download, with minimal metadata. It remains difficult to carry out basic research tasks such as searching across multiple repositories for documents mentioning particular named entities, annotating a relationship between a resource in one repository and one in another repository, or running an arbitrary Natural Language Processing algorithm on all the texts contained in one or more repositories. In the interest of transparency in research, it would be particularly valuable for researchers to be able to share new data that is partly based on existing data, in a way that preserves the links between the existing and new data (genesis of research data), and allows future researchers to follow those links to verify what was done, as well as to repeat the process by adding another layer of new results. But this remains a highly challenging task because of the lack of adequate infrastructure. Furthermore, it should be possible to do all this in a consistent manner from within a virtual research environment (VRE) of the researcher's choosing, so that she would not have to learn the different search capabilities, metadata formats, and access methods provided by a myriad of different repositories and data sets. But the great diversity of (often ad-hoc) technologies used to create these repositories makes it very difficult to construct any such VRE. While some repositories offer Application Programming Interfaces (APIs), the great diversity of these APIs means that it is highly burdensome to design software that works with multiple repositories. There is no common repository interface that would enable a VRE or other software application to work in a convenient, unified manner with different sorts of research data, produced by different research projects, in different disciplines, and stored in different repositories.

We argue that such functionality, which we would describe as the continuous availability and reusability of digital qualitative research data in the humanities, requires a common, flexible data representation and storage technology capable of performing queries across large quantities of heterogeneous data, organised according to project-specific data structures that cannot be known in advance. It also requires a convenient, storage-independent way for VREs and automated data-processing software to access, query, and add to this data. In our presentation, we introduce Knora (Knowledge Organization, Representation, and Annotation), a design and implementation aimed at solving both these problems.

To solve the data representation and storage problem, Knora describes a way of representing humanities data as RDF graphs, using OWL ontologies that express abstract, cross-disciplinary commonalities in the structure and semantics of research data. Each research project extends these abstractions by providing its own project-specific ontology that more specifically describes the structure and semantics of its data. It is possible to define OWL equivalences between these abstract entities and those used in other sorts of RDF-based repositories. Existing non-RDF repositories can readily be converted to an RDF format based on the proposed abstractions. This design makes it possible to preserve the semantics of data imported from relational databases, XML-based markup systems, and other types of storage, as well as to query, annotate, and link together heterogeneous data in a unified way. By proposing a shared, standards-based, extensible infrastructure for diverse humanities projects, we also address the task of conversion and migration caused by the obsolescence of file and data formats in an efficient and feasible manner.

To solve the access problem, Knora implements a generic HTTP-based API. This API allows applications to query and work with data in terms of the concepts expressed by the generic Knora ontologies, without dealing with the complexities of the underlying storage system and its query language (e.g. SPARQL). The Knora API also provides features that are not part of SPARQL, such as access control and automatic versioning of data. While the Knora API is best suited to interacting with RDF repositories based on the Knora ontologies, it can also be implemented as a gateway to other sorts of repositories, including non-RDF repositories.

Finally, we describe an implementation of this API, which we have used to implement a browser-based VRE; both the API and the VRE are being used by a number of research projects in different humanities disciplines. The VRE sends HTTP requests to the API, and receives responses in JSON format. We discuss lessons learned from the development and use of this implementation, and the resulting redesign and reimplementation based on these lessons. We conclude that this approach can greatly increase the value of research data long after the completion of the projects that produced the data, by placing subsequent researchers on an equal footing with the original researchers, enabling them to reuse, annotate, and extend it, and to combine it with data from other projects in a consistent way.

Vom digitalen Hausmeister zum international vernetzten Forschungsdatenzentrum

Die Entwicklung der digitalen Infrastruktur des Hamburger Zentrums für Sprachkorpora

Daniel Jettka [daniel.jettka@uni-hamburg.de], Universität Hamburg
Hanna Hedeland [hanna.hedeland@uni-hamburg.de], Universität Hamburg
Timm Lehmberg [timm.lehmberg@uni-hamburg.de], Universität Hamburg

Zusammenfassung

Der Beitrag stellt den aktuellen Entwicklungsstand der digitalen Infrastruktur des Hamburger Zentrums für Sprachkorpora (HZSK) dar und zeigt dessen Entwicklung vom datenverarbeitenden Teilprojekt eines Sonderforschungsbereichs hin zu einem unabhängigen, international vernetzten, zertifizierten Zentrum für linguistische Forschungsdaten und -dienste. Die Antwort auf die sich stark verändernden Anforderungen liegt hierbei in einer modularen, digitalen Architektur, die auf Basis eines Repositoriums für linguistische Ressourcen mit unterschiedlichen Nutzer- und Programmschnittstellen und durch die Vernetzung mit der Forschungsinfrastruktur CLARIN und anderen Forschungsprojekten eine solide und anpassungsfähige Grundlage für eine dauerhaft zugängliche und nutzbare Forschungsumgebung bildet.

Abstract

Das Hamburger Zentrum für Sprachkorpora (HZSK), hervorgegangen aus dem datenverarbeitenden Teilprojekt des 2011 beendeten Sonderforschungsbereichs 538 "Mehrsprachigkeit", besteht aus einem Verbund von Forschenden, die digitale Methoden in der linguistischen Forschung einsetzen und diese voranbringen wollen. Im Kern dieses Zusammenschlusses arbeiten zwei Forschungsprojekte, das Hamburger CLARIN-D-Teilprojekt und das DFG-Projekt "Etablierung eines Schwerpunkts Mehrsprachigkeit und gesprochene Sprache am HZSK", an der Weiterentwicklung einer digitalen Infrastruktur, die die am HZSK vorgehaltenen linguistischen Ressourcen dauerhaft verfügbar und nutzbar machen soll.

Mit der Entwicklung des EXMARaLDA-Systems während der Projektlaufzeit des SFB 538 wurde durch den institutionellen Vorgänger des HZSK der Grundstein für eine digitale Infrastruktur für Forschungsprojekte, die Daten auf der Grundlage gesprochener Sprache erheben und auswerten, gelegt. Während der Fokus hierbei zunächst auf der Unterstützung einzelner Forschender bzw. individuell arbeitender Forschungsprojekte durch die Entwicklung und Bereitstellung von Desktop-Software, Formaten und Workflows für die Transkription und Analyse gesprochener Sprache lag, sind durch die Ziele des über den SFB hinausgehend konstituierten HZSK (z.B. langfristige Verwendbarkeit und Verfügbarkeit digitaler Sprachdaten) und mit dem Entstehen überregionaler Forschungsinfrastrukturen die Publikation und Nachnutzung von Forschungsdaten sowie die Vernetzung mit anderen infrastrukturellen Einrichtungen und Diensten zunehmend in den Mittelpunkt gerückt.

Ein Hauptaugenmerk der Strategie zur langfristigen Erhaltung der Verfügbarkeit und Nutzbarkeit der Daten, Werkzeuge und Dienste am HZSK liegt demnach in der Vernetzung mit der CLARIN-Infrastruktur, durch die die Integration und Kompatibilität der Ressourcen gesichert werden kann, z.B. durch:

- die Unterstützung sich weiterentwickelnder Standards und Best Practices für die Aufbereitung und Vorhaltung linguistischer Daten,
- die gegenseitige Absicherung der Datenbestände der CLARIN-Zentren,
- die Bewältigung nichttechnischer Anforderungen (z.B. bzgl. rechtlicher Fragestellungen der Zugänglichkeit von Ressourcen, Lizenztypen und Nutzungsvereinbarungen),
- die Zertifizierung von CLARIN-Zentren in Bezug auf die Umsetzung der CLARIN-Standards, die auch die Erfüllung externer Standards, wie etwa der des Data Seal of Approval, erfordert.

Diese Rahmenbedingungen und Zielsetzungen wirken sich auf die Implementierung der lokalen Infrastruktur in an CLARIN-D angebundenen Forschungseinrichtungen aus:

- Standardisierte Metadaten zu den im Repositorium vorhandenen Forschungsdaten werden über eine OAI-PMH-Schnittstelle ausgeliefert, damit sie automatisch gesammelt werden können und die Forschungsdaten beispielsweise im CLARIN Virtual Language Observatory (VLO) sichtbar werden.
- Mit der CLARIN Federated Content Search (FCS) wurde eine zentrenübergreifende Suche in den verfügbaren Ressourcen geschaffen, bei der Suchanfragen von einem zentralen Einstiegspunkt an die jeweiligen Zentren über SRU/CQL beantwortet werden können.
- Ressourcen werden mit Hilfe von Persistenten Identifizierern (PID) auf Basis des Handle-Systems dauerhaft identifizierbar und referenzierbar gemacht, sodass sie beispielsweise in Forschungsarbeiten zuverlässig zitiert werden können.
- Der webbasierte Zugriff auf Ressourcen über ein Single Sign-on-Verfahren, bei dem Benutzerdaten von teilnehmenden Institutionen (Identity Providers) mit den zugangsbeschränkten Benutzerschnittstellen der Zentren (Service Providers) verknüpft und über das SAML-Protokoll übertragen werden, ermöglicht den föderierten Zugriff auf Forschungsdaten.

Diese CLARIN-spezifischen Schnittstellen und Module sind am HZSK an ein Repositorium angebunden, das auf der Flexible Extensible Digital Object Repository Architecture (Fedora) basiert und als integraler Bestandteil der digitalen Infrastruktur am HZSK angesehen werden kann. Fedora stellt hierbei eine weitverbreitete und vielversprechende Lösung zur langfristigen Speicherung und Verfügbarmachung von Forschungsdaten dar. Neben der genannten CLARIN-spezifischen Funktionalität beinhaltet die digitale Infrastruktur am HZSK weitere Merkmale bzw. Funktionen, die auf die speziellen Anforderungen des Zentrums und seiner Nutzer zugeschnitten sind.

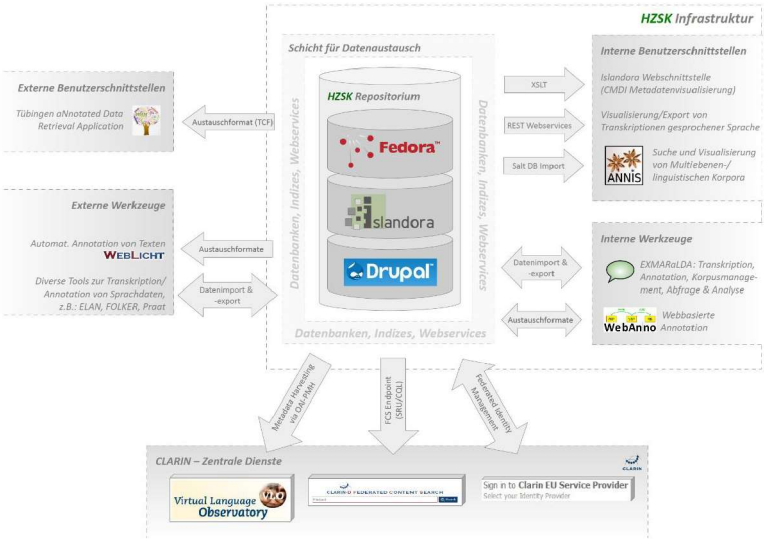
Da die vorgehaltenen Daten aufgrund von datenschutzrechtlichen Bestimmungen häufig nicht pauschal der gesamten wissenschaftlichen Öffentlichkeit zugänglich gemacht werden können, musste ein mit dem o.g. Single Sign-on kompatibles System zur Authentifizierung und Authorisierung von Nutzern entwickelt werden. Dieses basiert auf Erweiterungen des Content Management Systems Drupal (durch spezifische Module und das Software Framework Islandora), mit deren Hilfe Nutzer sich lokal oder per Single Sign-on anmelden können, um Zugang zu den Inhalten des Fedora-Repositoriums zu erhalten.

Zudem ist zu berücksichtigen, dass die Nutzer bei der Arbeit mit den Sprachkorpora oftmals explorativ-iterativ vorgehen und umfangreiche und detaillierte Informationen und

Recherchemöglichkeiten zu den vorhandenen Ressourcen benötigen. In die Benutzerschnittstelle des Repositoriums wurden RESTful Java Webservices integriert, die auf Anfrage spezifische Ansichten oder Exportformate der Transkriptionen liefern und dadurch die Weiterverarbeitung der Daten mit verschiedenen Werkzeugen sowie die Auswertung aus unterschiedlichen fachwissenschaftlichen Perspektiven unterstützen.

Während die webbasierte Benutzerschnittstelle zunächst vor allem auf Sprachkorpora zugeschnitten war, die mithilfe des EXMARaLDA-Systems erstellt wurden (beispielsweise durch eine weitgehende generische, einheitliche Visualisierung der Datensätze), wurde die Entwicklung der Infrastruktur, um die Aufnahme weiterer externer Ressourcen sowie die Anbindung zusätzlicher Werkzeuge und Dienste an die modularisierte Infrastruktur zu erleichtern, immer weiter von den EXMARaLDA-Formaten und Werkzeugen losgelöst und auf größere Flexibilität und Interoperabilität ausgerichtet. Gleichzeitig wurde die Kompatibilität der EXMARaLDA-Werkzeuge mit der CLARIN-D-Infrastruktur optimiert, sodass vorhandene EXMARaLDA-Daten zusätzlich in CLARIN-D-spezifische Formate exportiert oder direkt mit dem CLARIN-D-Dienst WebLicht, der eine Vielzahl linguistischer Werkzeuge als untereinander kombinierbare Webservices zur Verfügung stellt, bearbeitet werden können.

Durch die zunehmende Modularisierung der Infrastruktur wird die Anbindung weiterer Werkzeuge und Schnittstellen, wie beispielsweise WebAnno für webbasierte, kollaborative Annotation oder ANNIS für die Abfrage und Visualisierung von Korpusdaten, ermöglicht. Auf diese Weise kann auch künftig durch die Hinzufügung weiterer ressourcenspezifischer Schnittstellen und Werkzeuge in Form von zusätzlichen Modulen der Infrastruktur weiteren Nutzerbedürfnissen entsprochen werden.



Empirische Daten für die Kunstgeschichte erzeugen mit Artigo Social Image Tagging

Harald Klinke [h.klinke@lmu.de], Institut für Kunstgeschichte, Ludwig-Maximilians-Universität München

Zusammenfassung

Artigo is an online system that provides massive data on art-historical images. Thus, it provides invaluable additional information in order to find images, to sort and arrange them and by this producing derived data that serve as means for big data analysis. This paper shows the functions of the system, its generated data and the purpose for art history.

Abstract

Today, image databases in Art History still resemble digitized slide libraries. Jpg-files are associated with meta data just like the slide and its information on the frame. The fact that those images are present in a digital form, however, opens up to new possibilities. Massive tagging is one means to generate data that help finding images faster, analysing the database in new ways and developing new methods to reach scientific goals in Art History.

The Artigo system was developed by the department of History of Art in close collaboration with the department of Information Science at the University of Munich as an online "game with the purpose" to enrich art-historical image databases. It offers the user via a web browser a range of possibilities to enter keywords associated with images that are successively presented by the system to the user. The incentive for the user is to earn points according to the value of the entered keywords, to appear in a high score list and to play against an opponent live.

By this, to each image in the database, a great number of keywords are generated by humans – much faster (and cheaper) than any employed group of students could. The Artigo database has yielded already millions of tags. As images are for the computer only a matrix of pixels, those tags represent what is being seen in the image and help finding visual content in the database.

The benefit of those data, however, exceeds the mere finding of images by far. It opens the possibility of mining a large set of scientific data. Now it is possible to weave a semantic web among the images, to automatically cluster images by content or to statistically prove scientific hypotheses by means of big data analysis. This paper presents first results and takes a look into the future of digital methods in Art History by raising questions such as: What data does Art History need? What methods should we develop? And what is our relation to Information Science?

https://www.artigo.org

Database of historical places, persons, keywords, and lemmas

Natalia Korchagina [natalia.korchagina@ssrq-sds-fds.ch],
Schweizerische Rechtsquellenstiftung, Zürich, Switzerland

Summary

This article presents a non-relational database of historical persons, organizations, places, keywords, and lemmas. Initially planned as a tool for edition of the back-of-the-book indexes of the upcoming volumes of historical texts of Swiss Law Sources, it will also help to make our data available for the large public. The RDF graph data model provides means for knowledge discovery, allowing us to develop a mechanism for more than just database browsing.

Abstract

The importance of the representation of humanities material as structured, interconnected objects has grown with the recent emergence of Linked Open Data and the 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 data storage technologies, i.e. databases, is crucial for a digital humanities project.

The Swiss Law Sources Foundation has been publishing critical editions of Swiss historical legal texts for over a hundred years, for a total of 118 volumes. The collection contains texts in German, French, Italian, Romansh, and Latin. The texts' creation time ranges from the 10^{th} to the 18^{th} century, representing therefore a rich source not only of historical information, but also of diachronic linguistic data. Each of these volumes contains a back-of-the-book index of persons, places, keywords, and lemmas. Until recently, this index was composed manually for each new volume. Therefore, inconsistencies can often be found when comparing entries for the same entity in different volumes.

The database should facilitate the edition of the index of future volumes, as well as be a starting point for users looking for information on a specific person/place/keyword/lemma which could have been mentioned in one of the volumes. The database will have the four basic CRUD (create, read, update, delete) functions of persistent storage, and will provide its users with an intuitive GUI. To give some more details on a historical person or place the database entries will be linked whenever possible with the corresponding entries in international reference sources, e.g. GND (Integrated authority file of the German National Library), and Swiss reference sources, e.g. HLS (Historische Lexikon der Schweiz). Additionally, the database makes part of the planned Swiss metadata hub for historical entities, *histHub.net*.

The technology to be used is an RDF triple store, a NoSQL (non-relational) mechanism for storing and retrieval of data. In a triple store, each data entity is composed of a subject-predicate-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. Furthermore, an RDF triple store has a simple and uniform standard data model, used for data interchange in the emerging Semantic Web and Linked Open Data. As a future goal, we aim to integrate our RDF data into other international projects (e.g. DBpedia or Europeana) for a higher visibility.

A set of controlled vocabularies will be integrated into the database. Such units as first names or keywords (i.e. role names, professions, types of places and organizations) are used repeatedly in the source texts and in the index. It seems reasonable to store them in a consistent manner, in order to enable a more efficient querying and consistent updating of the database.

The method of editing of controlled vocabularies proposed by the Labeling System project (i3mainz - Institute for Spatial Information and Surveying Technology, Leibnitz Institute of European History in Mainz) complies with the nature of our database. The system adopts the concept of Linked Open Data implementing technologies and formats (e.g. standardized data models, such as SKOS and Dublin Core) to semantically model the relationship and meaning of concepts from a set of thesauri. The RDF graphs representing the vocabularies are accessible via a REST interface. The controlled vocabularies mentioned above will be developed in collaboration with the Labeling System project.

At the current stage of the project, a working prototype with a GUI is in production. The index of persons of the upcoming volume of the Swiss Law Sources is being edited with it. It contains about 127,000 triples describing 3400 persons and 1455 organizations. This data represents back-of-the-book indexes of persons and organizations of two volumes. The GUI has basic search possibilities, e.g. search by person's surname, forename and ID; as well as limited filtering, e.g. by the organization's type. The prototype is still in development, and many more querying functionalities are yet to come. At the end, we aim to develop a tool which can be more than just an interface for database browsing. Our goal is to enable knowledge discovery via complex queries, e.g. "Show me the number of persons who changed their place of residence in the 15^{th} century in a particular region". This can provide new insights into social and economical processes in the past, and can be interesting not only for historians.

INEL

Eine Infrastruktur zur Dokumentation indigener nordeurasischer Sprachen

Beata Wagner Nagy [beata.wagner-nagy@uni-hamburg.de], Universität Hamburg
Hanna Hedeland [hanna.hedeland@uni-hamburg.de], Universität Hamburg
Timm Lehmberg [timm.lehmberg@uni-hamburg.de], Universität Hamburg
Michael Rießler [michael.riessler@skandinavistik.uni-freiburg.de] Universität Freiburg

Zusammenfassung

Das in dem Vortrag vorgestellte Langzeitprojekt INEL (Grammatical Descriptions, Corpora, and Language Technology for Indigenous Northern Eurasian Languages), das am Institut für Finnougristik und Uralistik (IFUU[1]) und dem Hamburger Zentrum für Sprachkorpora (HZSK[2])) der Universität Hamburg verortet ist, zielt darauf ab, erstmalig die dringend erforderliche Erschließung der sprachlichen Ressourcen des genealogisch diversen nordeurasischen Sprachraums (s. Abb. 1 Der geographische Skopus des Projekts) zu leisten. Durch den Einsatz von State-of-the-Art-Methoden und -Werkzeugen der linguistischen Datenaufbereitung, die bisher nur für gut erforschten Sprachen und Varietäten zum Einsatz kamen, wird eine Lücke in diesen für die empirische Sprachwissenschaft bisher schlecht zugänglichen Arealen der Welt nachhaltig geschlossen. Dieses ehrgeizige Ziel stellt hohe Anforderungen an die Organisation von Arbeitsabläufen und erfordert zudem die Schaffung einer eigenen digitale Forschungsinfrastruktur sowie umfassende technische und internationale wissenschaftliche Vernetzung.

Ausgangslage

Aufgrund des drohenden Verfalls der zum großen Teil auf obsoleten Originalträgern (Wachswalzen, Schellackplatten, Mikrofiche etc.) in verschiedenen Archiven gelagerten Audio-Aufnahmen, Niederschriften und Beschreibungen, schließt sich in absehbarer Zeit das Fenster für einen Erhalt der Daten und somit auch für eine Überlieferung an kommende Generationen. Gleichzeitig gehen die Sprecherzahlen vieler Sprachen und Varietäten stetig zurück. Indem existierende Materialien zu digitalen Korpora aufbereitet und der bisherige Gesamtbestand um neue Ressourcen ergänzt wird, kann dieses Erbe als wertvolle empirische Basis für vielfältige Forschungsvorhaben erhalten werden. Vielmehr als nur ein digitales Archiv wird das Resultat dieses Projektes jedoch eine umfassende virtuelle Forschungsumgebung sein, die durch die Integration in supranationale Forschungsinfrastrukturen der wissenschaftlichen Öffentlichkeit dauerhaft zugänglich gemacht wird. Der inno-

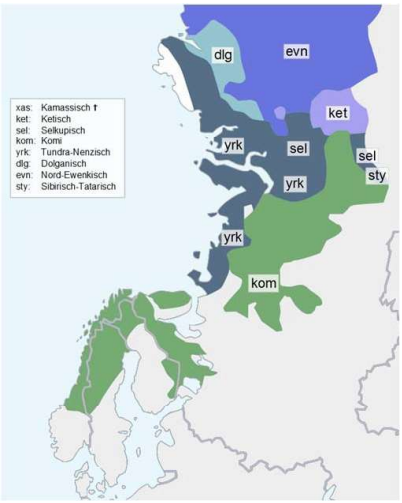


Abb. 1 Der geographische Skopus des Projekts

[1] https://www.slm.uni-hamburg.de/ifuu
[2] https://corpora.uni-hamburg.de/

vative Charakter des Projektes besteht zunächst darin, existierende Beschreibungen einzelner nordeurasischer Sprachen und Varietäten, die aufgrund der bisher begrenzten Auswahl von verfügbaren Sprechern und Genres eher partikuläre Idiolekte dokumentieren, zusammenzutragen und mit ergänzenden Korpora als umfangreiche digitale Ressource zugänglich zu machen. Durch die so geschaffene, der Vielfalt der Sprache angemessenen, Datenbasis werden für zukünftige Generationen von Forschenden erstmalig varietätenübergreifende Analysen möglich, etwa die Erforschung kontaktinduzierter Sprachveränderungen, Anwendungen aus dem Bereich der Dialektometrie oder sprachsoziologische Untersuchungen. Die unterschiedlichen Erhebungszeiten der Sprachdaten erlauben zudem erstmalig datengestützte Untersuchungen von diachronem Sprachwandel sowie Grammatikalisierungsprozessen. Ebenso bedeutend sind die Art des Zugangs zu den Sprachdaten und die damit verbundenen Analysemöglichkeiten. Die Sprachdaten können in der Forschungsumgebung kollaborativ und dezentral um beliebige weitere Beschreibungsebenen angereichert werden, die dann für verschiedene Auswertungsszenarien zur Verfügung stehen. Auf diese Weise wird die virtuelle Forschungsumgebung modular aufgebaut und in vielen Fällen so generisch sein, dass auch die Resultate technologischer und methodologischer Entwicklungen der akademischen Öffentlichkeit als Best Practices und als konkrete Grundlage für vergleichbare Vorhaben zur Verfügung stehen werden.

Arbeitspakete und Teilprojekte

Das Projekt gliedert sich in insgesamt zwölf Teilprojekte, von denen elf jeweils die Erschließung einer der zu erfassenden Sprachvarietäten zum Gegenstand haben. Das zwölfte technisch-infrastrukturelle Teilprojekt läuft im Gegensatz zu den elf jeweils auf drei Jahre angelegten Erschließungsprojekten durchgängig über die gesamte Projektlaufzeit und schafft somit die notwendige Kontinuität für die Entwicklung, Anpassung und Vermittlung der Funktionalitäten der technischen Infrastruktur. Das Arbeitsprogramm wird für die jeweiligen Teilprojekte aber auch teilprojektübergreifend in Form von methodischen Arbeitspaketen umgesetzt:

Arbeitspaket 1: Korpusaufbau
In diesem Arbeitspaket werden existierende Ressourcen erschlossen, kuratiert und aufbereitet sowie ggf. um neu zu akquirierende und zu erhebende Daten ergänzt. Die dabei erforderlichen Arbeitsschritte sind, wie in Abb. 2 dargestellt, modularisiert.

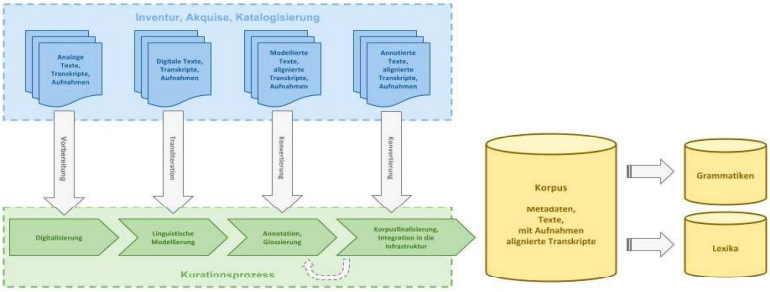


Abb. 2 Variabler modularisierter Workflow des Arbeitspaket 1

Die Module entsprechen jeweils Aufbereitungsschritten, die abhängig vom Ausgangszustand der einzelnen Ressource für die Integration in den Gesamtbestand erforderlich sind und werden im Rahmen des Vortrages ausführlich erläutert.

Arbeitspaket 2: Infrastruktur und Best Practices
Die zu errichtende Infrastruktur basiert in vielerlei Hinsicht auf den Vorarbeiten des

HZSK sowie generell auf vorangegangenen Erkenntnissen aus dem Aufbau digitaler Forschungsumgebungen. Große Teile der gewünschten Funktionalitäten wurden dabei bereits im Vorfeld in Form von Standalone-Werkzeugen oder als Komponenten in Desktop- oder Webapplikationen (bspw. EXMARaLDA[3]) oder der HZSK Infrastruktur in seiner Funktion als Zentrum CLARIN-D-Verbund entwickelt. Einen integralen Bestandteil der Infrastruktur bildet ein Repositorium, mit dessen Hilfe die nachhaltige Datenvorhaltung gewährleistet werden kann. Weitere Komponenten, die unmittelbar daran anknüpfen, bilden zusätzliche relevante Aspekte der gewünschten Funktionalität der Infrastruktur ab, wie etwa die kollaborative Bearbeitung und Aufbereitung von Daten in der Arbeitsumgebung, die Auslieferung von Metadaten an Kataloge und Archive, welche die Auffindbarkeit der Ressourcen für andere Forscher ermöglicht, sowie Schnittstellen für die Exploration und Analyse der vorgehaltenen Ressourcen (s. Abb. 3). Auch die fortlaufende Anbindung an bzw. Vernetzung mit weiteren bestehenden Forschungsinfrastrukturen wird als essentielles Merkmal des Projektes betrachtet.

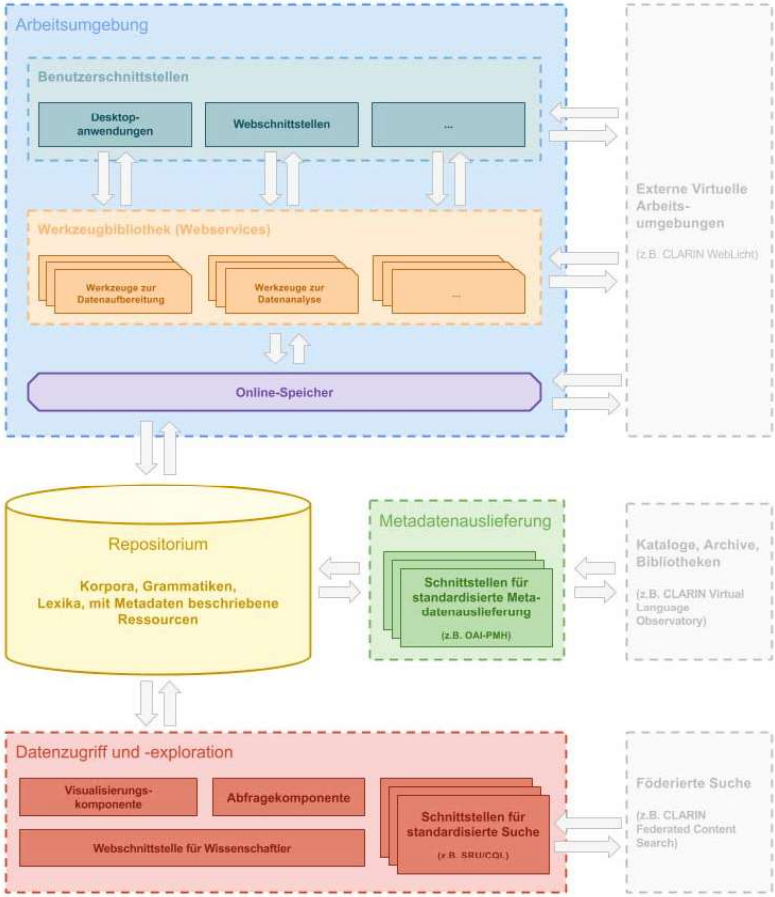


Abb. 3 Die Struktur der INEL Forschungsumgebung

[3] Extensible Markup Language for Discourse Annotation (http://www.exmaralda.org/)

Arbeitspaket 3: Evaluation und Dissemination:
Neben der organisatorischen und der technisch-infrastrukturellen Organisation und Vernetzung, die mit den Arbeitspaketen 1 und 2 abgedeckt ist, erfordert ein Langzeitprojekt wie INEL zudem umfassende Arbeiten im Bereich der Dissemination und den Austauschs im supranationalen interdisziplinären Kontexten mit anderen Forschenden. Um neue Arbeitsinstrumente zu entwickeln und zu diskutieren, Arbeitspläne zu koordinieren, aktuelle Forschungsfragen und die mit der Projektarbeit zusammenhängenden praktischen Fragestellungen zu diskutieren sowie zu Zwecken der Fortbildung werden im Rahmen von INEL jährliche Workshops abgehalten, an denen ausgewählte Konsultanten sowie Kooperationspartner aus dem Ausland und die Projektmitarbeiter selbst teilnehmen.

Die hier angerissenen Aspekte des Langzeitvorhabens INEL werden im Rahmen des Vortrages umfassend und anhand konkreter Beispiele erläutert und diskutiert.

Wissenstransfer und Wissensressourcen:

Support und Helpdesk in den Digital Humanities

Timm Lehmberg [timm.lehmberg@uni-hamburg.de], Universität Hamburg

Zusammenfassung

Ein vitales Support- und Helpdesk Angebot ist eine entscheidende Schnittstelle für den bidirektionalen Informationsfluss zwischen web- und zentrenbasierte Forschungsinfrastrukturen wie CLARIN oder DARIAH auf der einen und der Anwendercommunity auf der anderen Seite. Es gewährleistet die Sichtbarkeit und Dissemination der von der Infrastruktur zur Verfügung gestellten Dienste, Werkzeuge und Ressourcen; gleichzeitig garantiert es einen erfolgreichen Einsatz derselben und dient den Betreibern der Infrastruktur als Monitor für die Art und den Umfang der Nutzung der Infrastruktur und ihrer Komponenten sowie eventuellen Nutzerproblemen bis hin zu technischen Bugs.

Abstract

Um dieser Aufgabe gerecht zu werden, muss das Angebot zentralisiert sein, um die vielfältigen Expertisen und Wissensressourcen einer Forschungsinfrastruktur zu bündeln, zu kanalisieren und das Wissen an die Anwender weiterzureichen. Ebenso muss es dem dezentralen Prinzip der Forschungsinfrastruktur gerecht werden, indem lokale (zumeist zentrenspezifische) Supportroutinen für bestimmte Dienste oder Ressourcen integriert und Teil des zentralen Angebots werden, aber weiterhin autark existieren können. Hieraus resultieren hohe Anforderungen an die Konzeption von Supportworkflows und die technische Implementierung eines Support- und Helpdesk-Portals. Dieser Vortrag wird einen vollständigen Überblick über die konzeptionellen Überlegungen zur Schaffung des CLARIN-D Helpdesk, eine Beschreibung seiner technischen Implementierung mithilfe eines Ticketing Systems im Juni 2013 sowie einen Erfahrungsbericht über die bisherige Laufzeit des Portals liefern. Im Vordergrund stehen dabei Überlegungen zur Optimierung von Supportroutinen und die Schaffung und Integration der bei der Durchführung des Anwendersupports entstehenden Wissensressourcen.

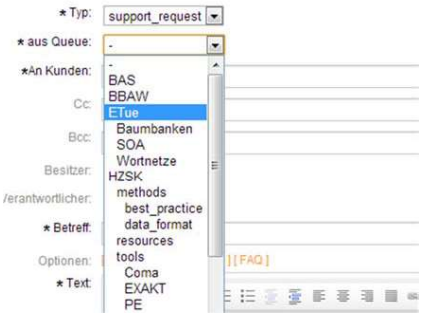


Abbildung 1 - Taxonomische Ordnung von Expertisen und Supportbereichen.

1. Strukturelle Vorüberlegungen

Das Grundprinzip der Strukturierung von Supportroutinen im CLARIN-D-Helpdesk ist eine taxonomische Gliederung von Wissensdomänen, Supportbereichen und den damit verbundenen Diensten und Ressourcen. Die Implementierung dieses Konzeptes erfolgt im Rahmen des CLARIN-D-Helpdesk (als Plattform dient das *Open Technology Real Services* (OTRS)[1] Ticketing System), durch ein hierarchisches Set so genannter Queues. Dabei handelt es sich um Containereinheiten zur Sortierung von Tickets und der Regelung von Zuständigkeiten innerhalb der Infrastruktur. Für die Praxis hat es sich als praktikabel erwiesen, *Queues* erster Ordnung einzelnen CLARIN-D-Zentren zuzuordnen, und in den jeweils untergeordneten Einheiten die zumeist zentrenspezifischen Supportbereiche einzugliedern.

[1] http://www.otrs.com/

Abbildung1 zeigt am Beispiel der Queue des CLARIN-D-Zentrums an der Universität Tübingen (ETUe) die untergeordneten Einheiten für *Baumbanken*, *Wortnetze* und *Service Oriented Architectures*. Die einzelnen Sub-Queues sind jeweils mindestens einem lokalen Mitarbeiter mit administrativer Funktion, sowie (insbesondere im Fall von zentrenübergreifenden Kompetenzbereichen) prinzipiell beliebig vielen weiteren Angehörigen anderer CLARIN-D-Zentren zugeordnet, die in der Funktion als *Helpdesk Agents* agieren und eingehende Anfragen bearbeiten und diese ggf. an externe Experten weiterleiten. Solche Gruppen von Helpdesk Agents besitzen jeweils exklusive Rechte auf die ihnen zugeordneten Queues, gleichzeitig ist das System so ausgereichtet, dass alle Supportbereiche für jeden Angehörigen des Helpdesk sichtbar und dokumentiert sind. Auf diese Weise wurden klare Zuständigkeiten bei gleichzeitig maximaler Transparenz und Durchlässigkeit geschaffen sodass Synergien zwischen Zentren in optimaler Weise genutzt werden können. Zudem sind alle jemals beantworteten Supportanfragen per Volltextsuche zugänglich, sodass in Verbindung mit den FAQ-Listen und Dokumentationen (s.u.) eine sich kontinuierlich entwickelnde Wissensressource entsteht, die als wertvolle Grundlage für spätere Anfragen dient.

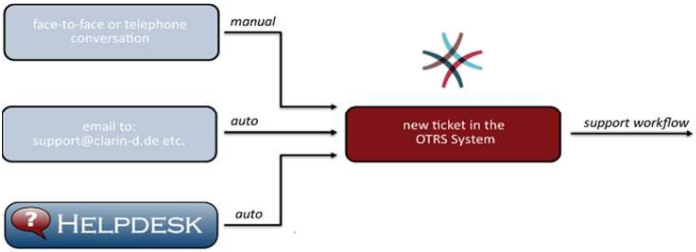


Abbildung 2 - Initiation von Support Workflows

2. Ticketing Workflows

Ein Supportvorgang (im weiteren Verlauf als Ticket bezeichnet), kann auf verschiedene Weise auf der Basis von Anwenderanfragen initiiert werden (siehe Abbildung 2).

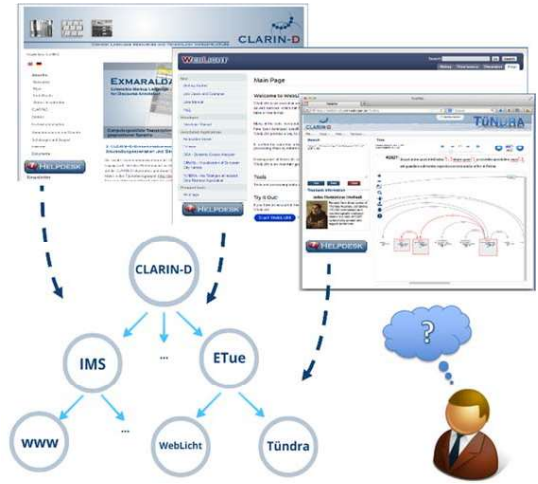


Abbildung 3 - Automatische Delegation von Tickets

Grundsätzlich können Tickets manuell von allen Angehörigen des Helpdesk angelegt werden. Eine automatische Generierung wiederum erfolgt durch die E-Mail eines Anwenders an eine zentrale oder domänen- und zentrenspezifische Adresse, die nach Bedarf in das System integriert werden. Zudem wurden gut erkennbare identifizierbare Call-to-Action-Buttons mit der Aufschrift „Helpdesk“ über das gesamte CLARIN-D-Webangebot verteilt, die ein Kontaktformular aufrufen, mit dessen Hilfe Anwender den Helpdesk konsultieren können. Abhängig von der gewählten E-Mail-Adresse bzw. der Position des aufgerufenen Helpdesk-Buttons, werden die resultierenden Tickets in entsprechende Queues einsortiert und einer zuständigen Gruppe von Helpdesk Agents zugeteilt. Der Anwender erhält darüber hinaus umgehend eine automatisch generierte Eingangsbestätigung die – in Abhängigkeit von der verwendeten E-Mail-Adresse oder Position des aufgerufenen Helpdesk Buttons im CLARIN-D-Webangebot – bereits Hinweise über anfragespezifische Wissensressourcen (bspw. FAQ-Listen, siehe 3) beinhalten kann. Das Ticket kann anschließend zwischen verschiedenen Angehörigen des Helpdesk delegiert, und diskutiert werden, auch können externe Experten über eine E-Mail-Schnittstelle hinzugezogen werden.

3. Integration von Wissensressourcen

Integraler Bestandteil des Ticketing Systems ist eine nach Kategorien unterteilte, sich in stetigem Ausbau befindliche Bibliothek von FAQ-Listen.[2] Diese werden teilweise in Verbindung mit existierenden Supportroutinen lokaler Zentren übernommen und integriert oder stammen von zentrenübergreifenden Portalen wie der CLARIN-D Legal Information Platform[3]. Gleichzeitig werden auf der Grundlage der kontinuierlichen Supportanfragen bestehende FAQ-Listen ergänzt, an technische Neuerungen angepasst und teilweise neu geschaffen. Den Helpdesk Agents stehen zudem zur Hilfe bei der Durchführung des Anwendersupports interne Versionen der FAQ-Listen zur Verfügung, die technische komplexere Inhalte bereitstellen. Eine wichtige Neuerung, mit deren Hilfe bessere Möglichkeiten der Integration des CLARIN-D-Helpdesk in dezentrale Webangebote geschaffen wurden, ist die Implementierung eines SOAP-basierten Webservice, mit dessen Hilfe das Angebot an FAQ-Listen durchsucht und abgerufen werden kann, sodass die Ressourcen künftig zentral gepflegt aber prinzipiell in jede beliebige Website eingebunden werden können und dort automatisch aktualisiert werden. Auf diese Weise wurde zudem Anbietern von Ressourcen und Werkzeugen für die Digital Humanities die Möglichkeit gegeben, Ihre FAQ-Listen und damit vernetzte Dokumentation in eine umfassenden supranationalen Infrastruktur einfließen zu lassen und von derselben profitieren, ohne auf lokale Präsenz des Angebotes verzichten zu müssen.

4. Stand und Ausblick

Seit Inbetriebnahme des CLARIN-D Support- und Helpdesk Portals und der Implementierung des Ticketing Systems im Juni 2013 ist die Zahl der Supportanfragen auf durchschnittlich 20 Tickets in der Woche (Stand April 2015) angestiegen. Gleichzeitig ist als Resultat der hier beschrieben Schritte der Optimierung von Ticketing Workflows sowie der Integration und Neuerstellung von Wissensressourcen der personelle Aufwand gesunken. Es ist damit zu rechnen, dass mit der Integration weiterer Portale und einem kontinuierlichen Anwachsen der Anwender der CLARIN-D-Infrastruktur mit einem weiteren Anstieg zu rechnen ist, dem auf diese Weise erfolgreich begegnet werden kann. Ein Fokus wird bei der künftigen Entwicklung auf der Implementierung weiterer Webservices zur Erstellung und Delegation von Tickets liegen, mit dessen Hilfe das Portal interoperabel mit anderen Ticketing Systemen werden wird. Dies wird beispielsweise einen nahtlosen Austausch von Informationen mit den von der CLARIN-D-Entwicklergemeinschaft genutzten Systemen oder Supportportalen anderer Forschungsinfrastrukturen (bspw. weiterer nationaler CLARIN-Konsortien) erlauben.

[2] http://support.clarin-d.de/otrs/public.pl
[3] http://de.clarin.eu/en/training-helpdesk/legal-helpdesk.html

Das Geschäftsmodell für den Betrieb der Virtuellen Forschungsumgebung FuD

Ergebnisbericht des DFG-Projektes „FuD2015“

Thomas Burch [burch@uni-trier.de], Universität Trier. Trier Center for Digital Humanities

Gisela Minn [minn@uni-trier.de], Universität Trier. Forschungszentrum Europa

Marina Lemaire* [marina.lemaire@uni-trier.de], Universität Trier. Servicezentrum eSciences

Yvonne Rommelfanger [rommelfanger@uni-trier.de], Universität Trier. Servicezentrum eSciences

Zusammenfassung

Das derzeitige Geschäftsmodell der Virtuellen Forschungsumgebung FuD wird vorgestellt, offene Fragen werden identifiziert sowie erste Erfahrungen bei der Umsetzung des Geschäftsmodells präsentiert. Für die Entwicklung eines Geschäftsmodells für den Regelbetrieb der Basisversion einer Virtuellen Forschungsumgebung (VFU) sowie für den Aufbau projektspezifischer Anwenderversionen und deren Betreuung wurde anhand verschiedener Forschungsprojekte eine detaillierte Prozessanalyse vorgenommen. Dabei wurden die notwendigen Aufgabenfelder definiert und die Zuständigkeit für die verschiedenen Arbeitsschritte den beteiligten Akteuren zugeordnet; die Workflows wurden fortlaufend evaluiert und angepasst. Ebenso wurden verschiedene Organisations- und Kostenmodelle anhand verschiedener Szenarien geprüft und dabei diverse rechtliche Bestimmungen (z.B. Haushalts- und Beihilferecht) bei diesen Planungen berücksichtigt.

Abstract

Von 2004 bis 2012 wurde die Virtuelle Forschungsumgebung FuD im Rahmen des Sonderforschungsbereiches 600 „Fremdheit und Armut“ in Zusammenarbeit mit dem Trier Center for Digital Humanities entwickelt. Das Softwaresystem unterstützt den gesamten Forschungsprozess von der Datenerhebung und -analyse, über die Redaktion und Publikation bis hin zur Archivierung. Es ermöglicht die zeit- und raumunabhängige, kollaborative Forschungsarbeit. Aufgrund der stetig größer werdenden Nutzergruppe und der Frage nach dem Fortbestehen der Software wurde ab 2013 im Rahmen des DFG-Projektes „FuD2015“ das Softwaresystem zu einer projektübergreifenden virtuellen Forschungsumgebung für die Geistes- und Sozialwissenschaften weiterentwickelt sowie ein Geschäftsmodell für deren Überführung in den Regelbetrieb entworfen und erprobt. Mit Beendigung des DFG-Projektes nimmt FuD an der Universität Trier im neugegründeten Servicezentrum eSciences den Regelbetrieb auf.

Der Vortrag stellt das gegenwärtige FuD-Geschäftsmodell vor. Es werden die drei Hauptfragen beantwortet: 1. Wie ist der Betrieb von FuD organisiert? 2. Welche Aufgaben haben die einzelnen Akteure? & 3. Welche Kosten fallen für den Betrieb von FuD an?

1. Der Regelbetrieb der FuD-Basisversion wird durch das *FuD-Basisteam*, das im Servicezentrum eSciences der Universität Trier arbeitet, sichergestellt. Es ist verantwortlich für die strategische Weiterentwicklung und die mit dem VFU-Regelbetrieb verbundenen

zentralen Aufgaben; ferner betreut und koordiniert es FuD-Anwenderinstanzen. Es ist zusammengesetzt aus Personal für Projektmanagement, IT-Koordination und Programmierung. Gemeinsam mit den FuD-Koordinationsteams der Forschungsinstitute werden der Releaseplan erstellt und die Programmierarbeiten koordiniert.

Mehrere FuD-Anwenderprojekte werden durch das *FuD-Koordinationsteam* betreut. Es handelt sich dabei um ein Team, das entweder an einer Universität bzw. Forschungseinrichtung für den lokalen Betrieb der VFU zuständig ist oder institutionenübergreifend projektspezifische FuD-Anwenderinstanzen mit gleicher oder ähnlicher Themenstellung koordiniert. Zur Zeit bestehen zwei Koordinationseinheiten: 1. Das Team des Trier Center for Digital Humanities, 2. FuD-Koordinatorin am Forschungszentrum Europa. Es ist zusammengesetzt aus Personal für IT-Koordination und Programmierung. Die Koordinationsteams organisieren die Zusammenarbeit verschiedener FuD-Anwenderprojekte, bestimmen deren gemeinsamen Entwicklungsbedarfe aufeinander ab und koordinieren das Anforderungs- und Releasemanagement mit dem FuD-Basisteam.

Das *FuD-Projektteam* ist zuständig für den Aufbau, die Anpassung und Weiterentwicklung einer projektspezifischen FuD-Instanz, die entweder von einer Universität bzw. Forschungseinrichtung bzw. mehreren Institutionen getragen wird; dieses Team ist in der Regel zusammengesetzt aus Fachwissenschaftlern und Personal für Programmierung, es ist auch möglich, dass das Projektteam ausschließlich aus Fachwissenschaftlern besteht, die notwendigen technischen Arbeiten für die FuD-Instanz werden in diesen Fällen vom zugehörigen FuD-Koordinationsteam oder vom -Basisteam übernommen.

2. Durch eine intensive Prozessanalyse, die die verschiedenen Betriebsszenarien der unterschiedlichsten FuD-Anwender berücksichtigte, ergaben sich folgende Aufgabenfelder, die sich bei der Entwicklung, Implementierung und Betreuung von FuD stellen: Anwendung, Community, Support, Training, Beratung und Management. Jedem Aufgabenfeld sind die notwendigen Arbeiten zugeordnet und die jeweils dafür zuständigen Akteure festgelegt. Diese systematische Zusammenstellung von Aufgabenbereichen, Akteuren und ihren Zuständigkeiten ist Grundlage für die Definition von Workflows und die Erstellung von Kostenmodellen. Dieses Basisschema für die Arbeitsorganisation dient der Qualitätssicherung und wird ständig überprüft, korrigiert, ergänzt und präzisiert mit dem Ziel, klar definierte Aufgabenfelder und Arbeitsabläufe festzulegen. Die Best-Practice-Workflows, Informations- und Präsentationsangebote, Planungshilfen und Mustervorlagen gewährleisten die effiziente Bearbeitung der Aufgaben über die Grenzen der verschiedenen FuD-Anwenderprojekte hinaus. Es handelt sich um Workflows für die Koordination des Anforderungs- und Releasemanagements, die Entwicklung und Implementierung neuer Softwarekomponenten, die Erarbeitung von Aufwands- und Kostenkalkulationen sowie die Angebotserstellung und Rechnungslegung, die Organisation der Zusammenarbeit zwischen den FuD-Teams mit Hilfe des Projektmanagementtools „Redmine“, die Erstellung von Tutorials, die Weiterführung der Anwenderdokumentation und Community-Informationen sowie die Erstellung der englischsprachigen FuD-Basisversion.

Für die definierten Geschäftsfelder und die ihnen zugeordneten Teilaufgaben werden pro Projekt die aufgewendeten Arbeitszeiten dokumentiert. Hierzu wurde mit dem Projektmanagement- und Dokumentationstool Redmine ein Ticketsystem aufgebaut, das in einem differenzierten Kategoriensystem die Geschäftsprozesse dokumentiert. Die detaillierte Zeiterfassung entspricht den Anforderungen der Abteilung für Finanzen der Universität Trier, um eine sachgerechte beihilfe- und steuerrechtliche Bewertung der Arbeiten vorzunehmen. Auf der Basis der Prozessanalyse wurden der Personalbedarf sowie das notwendige Anforderungs- und Qualifikationsprofil der Beschäftigten ermittelt, um den Regelbetrieb im Servicezentrum eSciences an der Universität Trier zu gewährleisten. Eingerichtet wurden zum 01.06.2015 vier Personalstellen für die Arbeitsbereiche Projektmanagement, IT-Koordination, Softwarenentwicklung und Webdesign.

3. Die Kostenmodelle berücksichtigen alle Leistungen, die für die Bereitstellung einer funktionsfähigen FuD-Customerinstanz und deren effizienten Einsatz im Forschungsprojekt notwendig sind. Dabei fallen Personalkosten im Rahmen der Leistungen der FuD-Teams sowie Dritter an, ebenso Sachkosten. Das Kostenangebot wird auf der Basis eines schriftlichen Anforderungsprofils der FuD-Anwender erstellt, es wird ergänzt durch einen Leistungskatalog, in dem FuD-Anwender und FuD-Team die notwendigen Anpassungs- und Weiterentwicklungsarbeiten für die projektspezifische FuD-Instanz dokumentieren.

Das Angebot gliedert sich in Kosten für die Installation und projektspezifische Konfiguration der FuD-Basisversion für die unterschiedlichen Teilkomponenten, Kosten für projektspezifische Weiterentwicklungen der FuD-Basisversion, Kosten für Schulungen, Support, die Bereitstellung von Hardware, die Bereitstellung und Konfiguration externer Tools und Werkzeuge sowie sonstige Dienstleistungen (z.B. Datenimport/-export).

Bei FuD-Anwendungen, die für eine externe Forschungseinrichtung bereitgestellt werden, wird die Dienstleistung gemäß dem EU-Beihilferecht nach Vollkosten berechnet, da eine wirtschaftliche Tätigkeit vorliegt. Eine nicht-wirtschaftliche Tätigkeit liegt nur bei Forschungskooperation mit der Universität Trier vor, die die Weiterentwicklung von FuD notwendig macht, um die definierten Forschungsziele zu erreichen.

Die Finanzierung des Regelbetriebs der FuD-Basisversion setzt sich zusammen aus der Grundfinanzierung der Betreiber und ihrer Anwenderprojekte, Drittmitteln (Projektmittel und Programmpauschale) sowie aus Erlösen aus Aufträgen Dritter.

Forschungsdaten in historischen Projekten oder die Quelle im Netz

Ein Werkstattbericht

Daniel Burckhardt [burckhardtd@geschichte.hu-berlin.de], Institiut für die Geschichte der deutschen Juden, Hamburg

Dr. Anna Menny [anna.menny@public.uni-hamburg.de], Institiut für die Geschichte der deutschen Juden, Hamburg

Zusammenfassung

Die Online-Quellenedition „Schlüsseldokumente zur deutsch-jüdischen Geschichte von der frühen Neuzeit bis in die Gegenwart“ versteht sich als ein innovatives Projekt zur praktischen Anwendung und Weiterentwicklung bestehender digitaler Infrastrukturen sowie zur Erprobung digitaler Publikationsformen. In dem Projekt werden Standardformate der Digitalisierung und Langzeitarchivierung mit einer nutzerfreundlichen Präsentationsumgebung verbunden. Der Zugang zu den Quellen und Texten wird über drei Wege möglich sein: über Themenkategorien, über eine Zeitleiste und über eine Karte. Die redaktionelle Einbettung der digitalisierten Dokumente zeichnet das Projekt in besonderem Maße aus und erhöht zugleich die Nutzerfreundlichkeit dieses digitalen Informationsangebotes zur deutsch-jüdischen Geschichte. Am Beispiel des Projektes soll über die wachsende Bedeutung von fachspezifischen Repositorien und Forschungsdaten reflektiert werden.

Abstract

Was sind Forschungsdaten für die jüdische Geschichte und wie können sie im digitalen Zeitalter für das Fach gewinnbringend und nachhaltig aufbereitet und präsentiert werden? Im Gegensatz zu den Natur- sowie den empirisch angelegten Sozialwissenschaften spielen Experimente, Messungen, Erhebungen oder Befragungen für die Mehrzahl historischer Arbeiten kaum eine Rolle. Nach wie vor bilden Text-, Bild- und audiovisuelle Quellen gegebenenfalls ergänzt um materielle Artefakte die Basis historischer Auswertungen und Analysen. Zwar gibt es in der historischen Demografie sowie im Bereich der vor allem in der Wirtschafts-, Sozial- und Umweltgeschichte verankerten historischen Statistik eine längere Tradition der systematisierenden Datenaufbereitung. Aufgrund der Forderung nach einem quellenkritischen Umgang sind aber derart vorstrukturierte Datensätze ohne Rückverknüpfung zu ihrer Quellenbasis für historische Fragestellungen weitgehend wertlos. Im digitalen Zeitalter stellt sich jedoch auch in den traditionellen Geschichtswissenschaften die Frage nach den Forschungsdaten ganz neu: Einerseits kann herkömmliches Quellenmaterial kommentiert und durch (Norm-)Daten und weitere Informationen ergänzt werden, andererseits ergeben sich neue zeit- und ortsunabhängige Präsentations- und Nutzungsmöglichkeiten.

Welche Quellen und Quellengattungen als relevant erachtet werden, hing stets sowohl von den individuellen Fragestellungen als auch den aktuellen Forschungstrends und Rahmenbedingungen ab. Ließ sich die Quelle aber im vordigitalen Zeitalter allenfalls in Form ihrer gedruckten Edition oder Reproduktion von ihrer bewahrenden Institution lösen, wird sie in ihrer digitalisierten Form mobil. Sie kann im Netz bereitgestellt und dort von den Besuchern der jeweiligen Website kommentiert und annotiert werden. Dabei werden ebenso wie bei der Auszeichnung der Quelle mit Normdaten oder georeferenzierten Daten zusätzliche Forschungsdaten erstellt. Im Gegensatz zu den

analogen Metadaten können die digital bereitstehenden Daten leicht für Forschungsfragen nutzbar gemacht werden, etwa im Hinblick auf Netzwerkanalysen. Kehrseite einer solchen Präsentation von Quellenmaterial im Internet ist die Flüchtigkeit und Zufälligkeit des Mediums. Die Herauslösung der Quelle aus der sie bewahrenden Institution kann zu einer Entkontextualisierung beitragen und die systematische Recherche jenseits der Freitextsuche bei Google erschweren. Hinzu kommt, dass eine Dauerhaftigkeit des Angebots bei derzeitigem technischen Stand nur selten gewährleistet ist.

Es ist daher an der Zeit, über die Frage der Forschungsdaten und ihrer angemessenen Präsentation neu nachzudenken. Die Online-Quellenedition „Schlüsseldokumente zur deutsch-jüdischen Geschichte von der frühen Neuzeit bis in die Gegenwart“ versteht sich als ein Versuch, zentrales Quellenmaterial aus einer Vielzahl verstreuter Institutionen mit weiteren für die Forschung relevanten Informationen zu verknüpfen und zugleich den Anspruch einer kommentierenden Edition zu erfüllen. Durch die systematische Verknüpfung mit Normdaten (Personen, Orte, Organisationen) soll einerseits die Auffindbarkeit der Quellen durch Dienste wie OAI und GND-Beacon gewährleistet werden. Verbreitete Standardformate (TEI, METS, MODS), die Einbindung in ein Konzept der Langfristarchivierung auf Basis der MyCoRe-Infrastruktur des RRZ Hamburg sowie persistente URLs gewährleisten die dauerhafte Zitierbarkeit. Eine nutzerfreundliche Präsentationsumgebung soll zudem eine Sichtbarkeit des Quellenportals über den engen Nutzerkreis fachspezifischer Repositorien hinaus ermöglichen.
Als elektronisches Projekt bleibt die Edition offen für Korrekturen und Erweiterungen. Durch Berücksichtigung von akzeptierten Format-Standards, „Best-Practice“-Richtlinien aber auch von Creative Commons-Lizenzen für die Kommentare und soweit möglich auch für die digitalisierten Quellen soll von Beginn an die Nachnutzung der Inhalte durch Drittanbieter unterstützt werden.
Ausgehend von einem Werkstattbericht beleuchtet der Vortrag die Frage der historischen Forschungsdaten von verschiedenen Seiten und stellt die disziplin-spezifischen Besonderheiten des Begriffs zur Diskussion.

Patrick Sahle, Simone Kronenwett, "Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner ‘Data Center for the Humanities’". LIBREAS. Library Ideas, 23 (2013). http://libreas.eu/ausgabe23/09sahle/
Christof Schöch, „Big? Smart? Clean? Messy? Data in the Humanities“. Journal for Digital Humanities, 2.3 (2013). http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/

Relationale Datenbanken und Webtechnologie im Dienst der Geolinguistik: eine Projektvorstellung

Markus Kunzmann [kunzmann@itg.uni-muenchen.de], Ludwig-Maximilians-Universität München

Abstract

In diesem Beitrag wird ein von der DFG gefördertes, datenbank- und webbasiertes Projekt aus dem Bereich der Digital Humanities vorgestellt, welches sich die selektive und analytische Erschließung der im Alpenraum gesprochenen Sprachen und Dialekte zur Aufgabe gemacht hat. Das untersuchte Areal ist einerseits sprachlich sowie politisch stark fragmentiert (drei Sprachfamilien und mehrere Einzelsprachen), andererseits aus ethnographischer Sicht und hinsichtlich der natürlichen Gegebenheiten über die Grenzen der Nationalstaaten hinweg vergleichsweise homogen. Die Bereiche, die im Zentrum des Projekts stehen, betreffen die für das alpine Gebiet relevanten Themen Natur, Kulturgeschichte und kulturelle Gegenwart. Das Ziel des Projektes ist ein mehrsprachiger, digitaler, interaktiver und fortlaufend erweiterbarer Online-Sprachatlas.

Zu diesem Zweck werden bereits erhoben Sprachdaten aus bestehenden Sprachatlanten und Wörterbüchern aufbereitet und digital zugänglich gemacht. Dabei werden die Daten vervollständigt, gebündelt, mehrdimensional aufbereitet und sprachgeschichtlich analysiert. Außerdem schließt das Vorhaben Lücken im Belegraster dieser älteren Datenbestände durch Neuerhebungen, welche mittels Crowdsourcing durchgeführt werden.

Die Konzeption basiert auf dem Zusammenspiel von Geolinguistik und moderner Medientechnologie. Die technische Realisierung hierfür erfolgt in Form eines Internet-Portals, dessen Herzstück von einer relationalen Datenbank gebildet wird und das drei Hauptfunktionen erfüllt: es dient

1) der Dokumentation der Projektdaten und umfasst die Möglichkeit der interaktiven georeferenzierten Kartographie,
2) der Datenerhebung durch Crowdsourcing sowie
3) der eigenen kollaborativen Weiterentwicklung.

Das vorgestellte Projekt arbeitet mit einer relationalen Datenbank (MySQL) und reiner Webtechnologie. Durch sein kollaboratives und integratives Design ermöglicht es die Zusammenarbeit des in Deutschland ansässigen Projektteams mit zahlreichen Partnern im In- und Ausland. Zusätzlich wird für die Datenneuerhebung die crowd eingebunden.

Drei verschiedene Module kommen bei der Umsetzung des Forschungsvorhabens zum Einsatz: der Datenbestand aus der MySQL-Datenbank, der Programmcode der Webschnittstelle des Projektportals samt zugehöriger Wordpress-Datenbank sowie diverse Mediendateien (Text, Foto, Video), die sich in der Medienbibliothek der Webschnittstelle befinden. Die drei Module bilden ein konsistentes Ganzes mit

wechselseitigen Verknüpfungen und Abhängigkeiten. Der jeweils aktuelle Status der ersten beiden Module wird in regelmäßigen Abständen in Gestalt einer elektronischen Kopie „eingefroren“ und somit versioniert.

Das Prinzip der Datennachhaltigkeit ist ein zentrales Anliegen des Projektes: Die Projektdaten werden so verwaltet, dass sie über einen möglichst langen Zeitraum les- und verwendbar bleiben (das zugrundliegenden Konzept ist letztlich auf eine Konservierung ohne zeitliche Befristung ausgerichtet); dazu werden mehrere Kopien der Daten bei verschiedenen Institutionen archiviert, u. a. werden Repositorien wie Clarin-D genützt. Die Projektdaten selbst sind ebenfalls Bestandteil der in regelmäßigen Abständen durchgeführten Versionierungen, die zugleich der Archivierung dienen.

Durch die browserorientierte Entwicklung der Forschungsumgebung, sind alle Daten jederzeit für jedermann und vor allem plattformunabhängig einsehbar. Durch die Möglichkeit der Einbettung in andere wissenschaftliche Kontexte, beschränkt sich der Nutzerkreis nicht nur auf eine sprachwissenschaftliche Klientel, sondern bezieht ebenso Historiker, Ethnologen und interessierte Laien mit ein. Aus informatischer Perspektive ermöglicht speziell das relationale Datenmodell zudem eine Nutzbarkeit der Daten über den Projektzeitraum hinaus und vor allem die Integration in erweiterten oder vollständig veränderten Kontext.

In gleicher Weise ist maximale Flexibilität hinsichtlich des Datenaustauschs durch das Schnittstellenkonzept gesichert. Vor dem Hintergrund klar definierter Referenzen bezüglich Datenstrukturierung und Zeichenkodierung (als Referenztranskriptionssystem dient das Internationale Phonetische Alphabet [IPA]) beinhaltet das Gesamtkonzept ein flexibles System geeigneter Konvertierungsroutinen für den Datenimport und -export (z. B. XML, CSV, Fließtext etc.). Neben der auf diese Weise erzielten Unabhängigkeit von bestehenden Kodierungen und Strukturierungen wird durch das Schnittstellenkonzept vor allem die Vergleichbarkeit der aus unterschiedlichsten Quellen stammenden Materialien gewährleistet.

IANUS als fachspezifisches Forschungsdatenzentrum für die Altertumswissenschaften in Deutschland

Maurice Heinrich [maurice.heinrich@dainst.de] Autor & Referent

Zusammenfassung:

Der Beitrag stellt verschiedene Facetten des Projektes "IANUS - Forschungsdatenzentrum Archäologie & Altertumswissenschaften" vor, das seit 2011 von der DFG gefördert und vom DAI in Berlin koordiniert wird. Zentrale Themen sind die digitale Langzeitarchivierung von altertumswissenschaftlichen Forschungsdaten, die Erhebungen zu den Bedürfnissen der Fach-Community, aber auch der praktische Umgang mit digitalen Daten sowie sich daraus ergebende Konsequenzen. Des Weiteren werden Angebote wie die IT-Empfehlungen für einen nachhaltigen Umgang mit digitalen Daten sowie weitere Dienstleistungen im Bereich des Forschungsdatenmanagements vorgestellt werden. Auch wird ein kurzer Einblick in das Betreibermodell gegeben.

Abstract:

Entsprechend der von der DFG formulierten „Regeln zur guten wissenschaftlichen Praxis“ sollen Digitale Forschungsdaten für längere Zeiträume aufbewahrt werden, damit sie auch in Zukunft nachgenutzt werden können. Dieser Forderung stehen – vor allem im Bereich der digitalen Geisteswissenschaften – zur Zeit noch die schlechte Zugänglichkeit, die Verteilung auf unterschiedliche Institutionen und die methodisch-technische Heterogenität der Daten entgegen. Dies gilt insbesondere auch für die Altertumswissenschaften, in denen die Datenmengen in starkem Maße wachsen und die sich an einer Schnittstelle zwischen Geschichts-, Geo-, Sozial- und Naturwissenschaften befinden. Diese Fächervielfalt trägt in hohem Maße dazu bei, dass die Art der Forschungsmethoden wie Ausgrabungen, Oberflächenbegehungen, Bauaufnahmen, Materialanalysen, DNA-Sequenzierungen, Vermessungen, Bohrkerne, Übersetzungen und Interpretationen von Schriftzeugnissen, textwissenschaftliche Analysen mannigfaltig sind und in der Konsequenz auch die Daten sehr heterogen sind. Bei einem Großteil dieser Methoden werden sie direkt digital erzeugt (digital born data), verarbeitet, visualisiert und gesichert.

Um dem bevorstehenden Verlust von einmaligen, nicht-reproduzierbaren und zumeist unpublizierten digitalen Forschungsrohdaten vorzubeugen und ihre Informationen der Wissenschaft langfristig zugänglich zu machen, sind disziplinenspezifische Zentren für die digitalen Forschungsdaten notwendig.

Da in Deutschland bislang eine zentrale auf die Archäologien und Altertumswissenschaften spezialisierte Adresse fehlt, wird mit „IANUS – Forschungsdatenzentrum für Archäologie und Altertumswissenschaften“ eine Daten- und Serviceinfrastruktur aufgebaut, die über die Grenzen von Bundesländern, Institutionen und Projektlaufzeiten hinweg IT-Dienstleistungen anbietet. Zentrales Anliegen ist es also, digitale Daten gemäß dem OAIS-Standard langfristig zu sammeln, zu katalogisieren, zu kuratieren und zu archivieren. Je nach rechtlichen Rahmenbedingungen der Datenproduzenten werden die Dateien online zur Nachnutzung bereitgestellt. Auf diese Weise wird die Nachprüfbarkeit von Ergebnissen erleichtert, eine langfristige Nachnutzung von existierenden Daten ermöglicht, der Austausch von Fachinhalten verbessert und die Sichtbarkeit der ForscherInnen durch persistente Zitationen erhöht.

Da einerseits viele aktuelle Fragen durch den vermehrten Einsatz elektronischer Systeme beantwortet werden und andererseits sich die Technologien rasant weiterentwickeln, erhöhen sich auch die IT bezogenen Anforderungen an AltertumswissenschaftlerInnen. Zunehmend werden Kenntnisse im Umgang mit gängigen Statistik-, GIS- und Datenbankprogrammen oder zur Strukturierung von Dokumenten und deren Erschließung mit Metadaten vorausgesetzt. Auch im Bereich der Projektunterstützung, Ausbildung, Schulung und Beratung ist es deshalb notwendig, über die verschiedenen Institutionen und Projekte hinweg eine vertrauenswürdige und kompetente Dienstleistungen anzubieten.

Etablierte und erfolgreiche Vorbilder für IANUS sind der *Archaeology Data Service in York/UK*[1] und das *e-depot for nederlandse archaeologie in Den Haag/NL*[2], mit denen ein enger fachlicher Austausch zu Fragen des Datenmanagements, zu Softwaresystemen, Metadatenstandards, Kosten- und Finanzmodells sowie der Lizensierung von Forschungsdaten besteht.

In dem Vortrag wird zunächst ein Überblick über die Ziele und den Stand des Projektes IANUS gegeben, welches durch und für die Fachcommunity in Deutschland koordiniert wird. Zuerst wird die Fach-Community kurz skizziert werden, um dann deren Kenntnisse und Bedürfnisse vorzustellen und zu analysieren. Dazu wird es einen dezidierten Einblick in die Ergebnisse der Stakeholderanalyse geben, welche 2013 durchgeführt wurde und 2014-2015 detailliert ausgewertet und veröffentlicht wurden. Aus den Ergebnissen resultieren die zukünftigen Dienstleistungen welche IANUS anbieten wird. Auch die Ergebnisse aus einem Testbed das 2014 durchgeführt wurde, in welchem exemplarisch die Daten aus 20 altertumswissenschaftlichen Projekten analysiert wurden, um einen praktischen Eindruck von der Komplexität aber vor allem der Heterogenität der archäologischen Forschungsdaten zu bekommen, wird thematisiert werden.

Den Schwerpunkt der Präsentation bilden die Ergebnisse der beiden Erhebungen sowie deren Konsequenzen für die Langzeitarchivierung von altertumswissenschaftlichen Forschungsdaten und den daraus abzuleitenden Workflows. Ein einfaches Modell zum Prozess der Sicherstellung der Langzeitverfügbarkeit - von der Abgabe über die Kuratierung bis hin zur Bereitstellung der Forschungsdaten - wird vorgestellt werden. Abschließend wird ein Ausblick über das Finanzierungsmodell des zukünftigen Forschungsdatenzentrums gegeben werden.

[1] http://archaeologydataservice.ac.uk/

[2] http://www.edna.nl/

Überlegungen zu einem Betriebsmodell für etablierte wissenschaftliche Software

Thomas Schmidt [thomas.schmidt@ids-mannheim.de], Institut für Deutsche Sprache, Mannheim
Kai Wörner [kai.woerner@uni-hamburg.de], Universität Hamburg

Zusammenfassung

Abstract

Dieser Beitrag zeichnet den Weg des wissenschaftlichen Softwarepakets EXMARaLDA (Schmidt/Wörner 2014) von der Erstkonzeption über seine Entwicklung und Verbreitung bis heute nach, wo das Entwicklerteam den stetig wachsenden Anforderungen an Pflege, Weiterentwicklung und Support nicht mehr im Rahmen der bisherigen Modalitäten gerecht werden kann. Vor diesem Hintergrund werden Überlegungen zu einem Betriebs- oder Geschäftsmodell angestellt, das für mehr Nachhaltigkeit in der Pflege und Entwicklung etablierter wissenschaftlicher Software sorgen kann.

EXMARaLDA ist eine Sammlung von Werkzeugen zur Arbeit mit audiovisuellen („mündlichen“) Sprachkorpora. Die Entwicklung war zwischen 2000 und 2011 in ein drittmittelfinanziertes Projekt des SFB 538 Mehrsprachigkeit an der Universität Hamburg eingebunden. EXMARaLDA wurde innerhalb des SFB umfänglich für die empirische Arbeit eingesetzt und fand auch außerhalb des SFB schnell größere Verbreitung. Es dient als technische Basis mehrerer großer Referenzkorpora (z.B. für das Deutsche: SiN, Kellner/Lehmberg/Schröder/Wörner 2008, GeWiss, Fandrych/Meißner/Slavcheva 2014, KidKo, Rehbein/Schalowski/Wiese 2014, FOLK, Schmidt 2014) wird von vielen Forschungsprojekten zur Datenbearbeitung und -analyse eingesetzt und ist auch in der universitären Lehre weit verbreitet. Die monatlichen Downloadzahlen liegen zwischen 500 und 1000, das Entwicklerteam bearbeitet jährlich etwa 300 Supportanfragen. EXMARaLDA kann demnach als ein erfolgreiches und weit verbreitetes wissenschaftliches Softwarepaket gelten.

Nach Abschluss des SFB 538 wurde eine Anschlussfinanzierung für zwei Drittmittelprojekte im Bereich digitaler Forschungsinfrastrukturen eingeworben, die aber keine Mittel für Weiterentwicklung und Pflege der Software oder für Supportleistungen beinhalten. Die beiden Hauptentwickler sind mittlerweile in Positionen an Forschungsdatenzentren beschäftigt, in denen die Weiterentwicklung von EXMARaLDA gleichfalls keine vorrangige Rolle spielt.

In dieser Konstellation sind weder eine nachhaltige Pflege, noch ein Nutzersupport auf dem bisherigen Niveau, geschweige denn einige wünschenswerte Weiterentwicklungen des Systems oder ein aktives „Marketing“ für seine weitere Verbreitung zu leisten. Das bisherige „Betriebsmodell“ – Finanzierung von Entwicklung und Pflege der Software durch selbst eingeworbene Forschungsgelder, kostenlose Bereitstellung von Software und Support – muss durch ein Modell abgelöst werden, das dem aktuellen Entwicklungsstand Rechnung trägt. Dabei sind folgende Beobachtungen wichtig:

1. Die eigentliche Entwicklung des Systems nimmt stetig weniger Raum ein, da zum einen die Grundfunktionalität seit langem etabliert ist, zum anderen die Umsetzung kleinerer Feature Requests mit wachsender Erfahrung der Entwickler weniger Zeit in Anspruch nimmt.
2. Aufwändig bleiben Anpassungen an neue technische Umgebungen, etwa bei der Einführung neuer Betriebssysteme (z.B. kürzlich Windows 8 und MAC OS

Yosemite) oder neuer technischer Standards (z.B. kürzlich MPEG-4 für Videos). Bleiben solche Anpassungen aus, wächst der Supportbedarf.
3. Es wächst die Zahl anspruchsvollerer Feature Requests (z.B. Integration von Webdiensten zur Annotation), die überhaupt nicht mehr umgesetzt werden können, weil solche fortgeschrittene Funktionalität in der Entwicklung aufwändig, oft aber nur für einen kleinen Kreis von Nutzern interessant ist. Kosten-Nutzen-Abwägungen führen oft dazu, dass diese Arten von Erweiterung nicht implementiert werden.
4. Mit dem wachsenden und sich diversifizierenden Nutzerkreis müssten hingegen die Investitionen in Dokumentation, Support und Schulung für die Software stetig steigen. Beispielsweise haben steigende Nutzerzahlen außerhalb Deutschlands eine englische Übersetzung der Dokumentation notwendig gemacht, und die wachsende Sensibilisierung von Forschern für Fragen des Datenmanagements führt dazu, dass immer mehr Projekte nach technischer Beratung oder Schulung anfragen.

Um die fehlenden Kapazitäten zu schaffen, wären schnell und flexibel einsetzbare finanzielle Mittel erforderlich, die mit der Nachfrage nach Leistungen wachsen und nach aktueller „Auftragslage“ im höheren fünfstelligen Bereich liegen müssten. Diese Mittel müssten zum kleineren Teil eingesetzt werden, um die Entwickler des Systems soweit von anderen Aufgaben zu entlasten, dass eine kontinuierliche Betreuung des Systems (wieder) möglich wird, zum größeren Teil für qualifiziertes Personal, das technische Dokumentation erstellen und pflegen, Schulungen planen und durchführen, sowie klar definierte Software-Entwicklungsaufträge ausführen kann.

Es scheint kaum realistisch, diese Mittel über die Beantragung eigener Forschungsgelder oder aus den Haushalten der beteiligten akademischen Institutionen zu schaffen. Eine kostenpflichtige Lizenzierung der Software ist grundsätzlich denkbar, wegen des zu erwartenden Mehraufwandes und der geringen Akzeptanz kostenpflichtiger Software aber wenig erfolgversprechend.

Angestrebt wird daher ein Betriebsmodell, in dem die Software weiter kostenlos zur Verfügung gestellt wird, es aber Möglichkeiten gibt, kostenpflichtige Vereinbarungen mit dem Entwicklerteam einzugehen, die mit gewissen Leistungen oder Privilegien verbunden sind. Vorstellbar sind:

- Support-Verträge für Fachbereiche, die EXMARaLDA regelmäßig in der Lehre einsetzen;
- Support-Verträge für Forschungsprojekte, die ihre Datengrundlage mit EXMARaLDA bearbeiten;
- Support-Verträge für einzelne Studierende und Promovierende;
- Kostenpflichtige Schulungen für Lehrende, Forscherteams oder im Rahmen universitärer Lehrveranstaltungen;
- Kostenpflichtige Unterstützung bei der Aufbereitung (Kuration) existierender Datenbestände für die Arbeit mit EXMARaLDA;
- Partnerschaften mit Forschungsprojekten, die für ihre Arbeit spezielle Erweiterungen benötigen;
- Partnerschaften mit Fachverbänden u.Ä., deren Mitglieder ein Interesse an der Weiterentwicklung von EXMARaLDA haben;
- Micro-Payment auf freiwilliger Basis für einzelne Support-Fälle;
- Crowd-Funding für die Umsetzung aufwändiger Features, die für einen größeren Nutzerkreis von Interesse sind.

Ein solches Modell würde die Nutzer der Systems in plausibler und transparenter Weise an den Kosten für Entwicklung bzw. Support beteiligen, dafür sorgen, dass benötigte Mittel bedarfsgerecht geschaffen und eingesetzt werden und gleichzeitig sicherstellen,

dass weiterhin alle Nutzer von einer solchermaßen finanzierten Weiterentwicklung des Systems profitieren. Für die Umsetzung des Modells müssten eine geeignete Rechtsform, Mittel für die initiale Konzeption und Implementierung, sowie eine geeignete Kommunikationsstrategie gegenüber der wissenschaftlichen Community gefunden werden.

Referenzen

Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hgg.) (2014): Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen. Heidelberg: Synchron-Verlag. (= Wissenschaftskommunikation).

Kellner, Birgit / Lehmberg, Timm / Schröder, Ingrid / Wörner, Kai (2008): Data structures for the analysis of regional language variation. In: Storrer, Angelika / Geyken, Alexander / Siebert, Alexander / Würzner, Kay-Michael (Hrsg.): Text Resources and Lexical Knowledge. Berlin, S. 53-63. [Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008]

Rehbein, Ines; Schalowski, Sören, & Wiese, Heike (2014). The KiezDeutsch Korpus (KiDKo) Release 1.0. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC),May 24-31, 2014. Reykjavik, Iceland.

Schmidt, Thomas (2014): The Research and Teaching Corpus of Spoken German – FOLK. In: Proceedings of the Ninth confer-ence on International Language Resources and Evaluation (LREC'14), Reykjavik, Iceland: European Language Resources Association (ELRA).

Schmidt, Thomas & Wörner, Kai (2014): EXMARaLDA. In: Durand, Jacques / Gut, Ulrike / Kristofferson, Gjert (eds.): Oxford Handbook of Corpus Phonology, Oxford: University Press.

Vom analogen Archiv zum digitalen Forschungsdatenzentrum

Aktuelle Herausforderungen im Archiv für Gesprochenes Deutsch (AGD)

Thomas Schmidt [thomas.schmidt@ids-mannheim.de], Institut für Deutsche Sprache, Mannheim

Zusammenfassung

Dieser Beitrag diskutiert aktuelle Herausforderungen, die sich in der Arbeit des Archivs für Gesprochenes Deutsch (AGD) im Kontext sich entwickelnder digitaler Forschungsinfrastrukturen stellen, und beschreibt, welche Ansätze zu ihrer Bearbeitung derzeit verfolgt werden.

Abstract

Dieser Beitrag beschäftigt sich mit aktuellen Herausforderungen in der Arbeit des Archivs für Gesprochenes Deutsch (AGD, https://agd.ids-mannheim.de). Das AGD (bzw. seine Vorgängereinrichtung, das Deutsche Spracharchiv, DSAv) fungiert seit über 80 Jahren als zentrale Sammelstelle für Daten des gesprochenen Deutsch (Stift/Schmidt 2014). Derzeit ist es im Kontext sich entwickelnder digitaler Forschungsinfrastrukturen einem grundlegenden Wandel unterzogen.

Das Archiv wurde in den 1930er Jahren gegründet und hat bereits in den 1950er Jahren Aufgaben übernommen, die auch heute in den Zuständigkeitsbereich von Forschungsdatenzentren fallen. Dazu gehören insbesondere die Koordination und Durchführung größerer Datenerhebungen, die Übernahme von Datenbeständen aus externen Projekten, die Erschließung und Dokumentation der betreffenden Korpora, sowie deren dauerhafte Archivierung und Bereitstellung für die Nutzung durch die wissenschaftliche Community. Ende der 1960er Jahre wurde das Archiv an das Institut für Deutsche Sprache angeschlossen und ist seitdem eng an dessen Forschungsprogramm angebunden. Bis zum Beginn der 1990er Jahre erfolgten Datenhaltung, -dokumentation und -weitergabe ausschließlich in analoger Form. Die dann begonnene Digitalisierung der Bestände ist nun weitestgehend abgeschlossen; neu hinzukommende Daten werden in der Regel direkt in digitaler Form erhoben. Seit der Jahrtausendwende hat sich das Archiv in mehreren Projekten mit der Entwicklung von geeigneten Methoden der digitalen Erschließung und Distribution seiner Bestände befasst; seit etwa fünf Jahren sind die Aktivitäten des Archivs verstärkt in nationale und internationale Initiativen zum Aufbau digitaler Infrastrukturen (insb. CLARIN) eingebunden.

Während also Gegenstandsbereich und Aufgaben des Archivs über mehrere Jahrzehnte weitestgehend gleichgeblieben sind, haben sich die konkreten Bedingungen des Umgangs mit den Archivdaten und die Position des Archivs in der Forschungslandschaft in den letzten zwanzig Jahren grundlegend und mit zunehmender Geschwindigkeit gewandelt. Dieser Wandel bringt fundamental neue Möglichkeiten und Herausforderungen mit sich, von denen einige in diesem Beitrag diskutiert werden sollen:

- Die Reichweite des Archivs hat durch die Möglichkeit der digitalen Distribution von Daten eine andere Größenordnung erlangt. Solange Daten nur auf analogen Trägern vorlagen, war ihre Weitergabe mit großem Aufwand und hohen Kosten verbunden. Vor der Digitalisierung des Archivs wurden daher jährlich in der Regel nicht mehr als 20 oder 30 Anfragen zur Datenweitergabe bearbeitet. Für das neueste digitale

Angebot des Archivs, die Datenbank für Gesprochenes Deutsch (DGD, http://dgd.ids-mannheim.de, Schmidt 2014) beläuft sich die Zahl der jährlichen Neuregistrierungen hingegen auf über 1000. Die Vielfalt der Nutzer, die die Daten für unterschiedlichste Zwecke in Studium, Lehre oder Forschung verwenden, stellt das Archiv vor gänzlich neue Fragen zu Einsatzbereich und Nutzbarkeit seines Bestandes, die es bislang nur ansatzweise beantworten kann. Systematische Nutzerstudien und ein den neuen Bedingungen angepasstes Verständnis von Archiv-Dienstleistungen müssen entwickelt werden, um das Archiv dauerhaft auf diese vergrößerte Reichweite einzustellen.

- Mit der größeren Reichweite und den neuen Zugriffsmöglichkeiten wachsen auch die Ansprüche an die Qualität der Daten. Jede neue Betrachtungsweise und jeder neue Betrachter der Daten legen bisher unentdeckte Fehler oder Lücken in den Daten offen, viele Nutzer haben viele Wünsche für ihre Arbeit mit den Daten, und das Archiv ist mit der Frage konfrontiert, wie es dem Korrektur- und Erweiterungsbedarf Rechnung tragen und gleichzeitig die Stabilität der Daten gewährleisten kann. In diesem Zusammenhang müssen professionelle Methoden des Versionsmanagement entwickelt werden. Gleichzeitig eröffnen Crowdsourcing-Verfahren ganz neue Möglichkeiten, die große Zahl der Nutzer aktiv in den Prozess des Datenmanagements und der Qualitätssicherung einzubinden. Auch diese Aspekte sind bislang nur im Ansatz erkundet worden.
- Die Weiterentwicklung des digitalen Angebots des Archivs erfordert Kompetenzen im Software-Engineering und im Datenmanagement, die weder über klassische Ausbildungswege zum Archivar oder Dokumentar noch im Rahmen IT-naher Studiengänge vermittelt werden. Ob und wie einmal identifiziertem Entwicklungsbedarf nachgekommen werden kann, ist daher oft von personellen Zufällen abhängig und entsprechend schlecht planbar. Mittelfristig wird der Weg des Archivs zu einem digitalen Forschungsdatenzentrum daher auch davon abhängen, wie es ihm gelingt, auf solche Kompetenzen zuzugreifen.
- Schließlich gewinnt mit dem Aufbau von digitalen Forschungsinfrastrukturen auch die Vernetzung des Archivs mit vergleichbaren Einrichtungen an Bedeutung. Aus Nutzersicht sind beispielsweise der Zugriff auf standortübergreifende Datenkataloge oder die Möglichkeit föderierter Abfragen naheliegende neue Anforderungen. Aus Sicht des Archivs bietet eine stärkere Vernetzung nicht zuletzt Chancen der Arbeitsteilung, die z.B. das im vorigen Punkt angesprochene Problem der fehlenden Kompetenzen in manchen Bereichen lösen können. Diese Chancen werden sich aber nur dann nutzen lassen, wenn in allen relevanten Bereichen stabile und allgemein akzeptierte Standards entwickelt werden.

Der Beitrag wird die genannten Möglichkeiten und Herausforderungen näher beschreiben und vorstellen, welche Ansätze zu ihrer Bearbeitung das AGD derzeit verfolgt.

Referenzen

Schmidt, Thomas (2014): The Database for Spoken German – DGD2. In: Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC'14), Reykjavik, Iceland: European Language Resources Association (ELRA).

Stift, Ulf-Michael & Schmidt, Thomas (2014): Das Archiv für Gesprochenes Deutsch. In: Institut für Deutsche Sprache (ed.): Ansichten und Einsichten : 50 Jahre Institut für Deutsche Sprache, 360-375.

histHub.net

Ein mehrsprachiger Metadatenhub für Linked Open Data zu historischen Entitäten aller Epochen und Regionen der Schweiz

Pascale Sutter [pascale.sutter@ssrq-sds-fds.ch], Rechtsquellenstiftung des Schweizerischen Juristenvereins, Zürich, Mitglied des Konsortiums *histHub.net*

Zusammenfassung

Histhub.net ist ein Metadatenhub für Linked Open Data zu historischen Entitäten aller Epochen und Regionen der Schweiz: Erschliessung von Daten zu historischen Personen, Organisationen, Sachen und Geographika dank verschiedener Zugriffsformen und Entwicklung von kontrollierten Vokabularen in drei Landessprachen sowie multilaterale Verlinkung der Daten inkl. Anreicherung mit Normdaten

Abstract

Normdaten und strukturierte Datenerfassung sind grundlegende Elemente datenbank- und korporaübergreifender Erschliessung und Vernetzung. Dies ist in Expertenkreisen unbestritten, aber das Wissen um Nutzen und beste Vorgehensweise ist in der geisteswissenschaftlichen Forschung – auch in den Digital Humanities – noch nicht weit verbreitet. Das geplante Projekt *histHub.net* wird als kooperative Publikations- und Anlaufstelle für Normdaten und strukturierte Datenerfassung hier einen wesentlichen Beitrag leisten, damit die entsprechenden Techniken und offenen Standards rechtzeitig und effizient in den Prozess der digitalen Forschungsproduktion integriert werden.

Mit dem Kooperationsprojekt *histHub.net*, das von einem Konsortium – bestehend aus der Rechtsquellenstiftung des Schweizerischen Juristenvereins mit ihrer Sammlung Schweizerischer Rechtsquellen (SSRQ),[1] dem Historischen Lexikon der Schweiz (HLS),[2] den Diplomatischen Dokumenten der Schweiz (DDS),[3] dem Schweizerdeutschen Wörterbuch (Idiotikon)[4] und dem Portal der schweizerischen Ortsnamensforschung (ortsnamen.ch)[5] – realisiert wird, entsteht ein Metadatenhub für Linked Open Data zu historischen Entitäten aller Epochen und Regionen der Schweiz. Hauptziel des Projekts ist der Aufbau eines mehrsprachigen, modular angelegten, transnational ausgerichteten Metadatenhubs zur konsistenten Pflege und nachhaltigen Nutzung von normierten Metadaten aus den Geisteswissenschaften. *histHub.net* wird unterschiedlichste Forschungsressourcen vernetzen und langfristig erschliessen.

Die Metadaten zu historischen Personen, Sachen und Geographika werden dank der gemeinsamen Erarbeitung von mehrsprachigen Keywordlisten, Normdateien und kontrollierten Vokabularen zu historischen Ämtern/Berufen/Funktionen und Ortstypen mit Hilfe des vom Leibniz-Institut für Europäische Geschichte (IEG)[6] und dem Institut für raumbezogene Informations- und Messtechnik, Hochschule Mainz (I3 Mainz)[7] für

1 Vgl. www.ssrq-sds-fds.ch
2 Vgl. www.hls-dhs-dss.ch
3 Vgl. www.dodis.ch
4 Vgl. www.idiotikon.ch
5 Vgl. www.ortsnamen.ch
6 Vgl. www.ieg-mainz.de
7 Vgl. http://i3mainz.hs-mainz.de/

diese Aufgaben entwickelten Open-Source-Werkzeugs «Labeling System»[8], das die Verlinkung mit verschiedenen Referenzthesauri erlaubt, qualitativ hochstehend erschlossen. Die multilaterale Verlinkung der Daten der Projektpartner inkl. Anreicherung mit internationalen Normdaten (z. B. GND, VIAF, TGN, GeoNames) wird neue Zugänge und Fragestellungen verschiedenster Forschungsrichtungen ermöglichen. Auf *histHub.net* wird neben den herkömmlichen Abfragemasken das historisch-geographische Informationssystem Geovistory[9] der Visualisierung und Analyse der Daten in Zeit und Raum dienen. Die für kollaboratives Arbeiten offene Projektanlage soll die Langzeitwirkung sichern. Dank der angestrebten Mehrsprachigkeit positioniert sich *histHub.net* zudem im Kontext der internationalen Digital Humanities als Brückenkopf über die Sprachgrenzen hinweg.

Bezüglich europäischer Forschungsinfrastrukturinitiativen steht *histHub.net* nicht in Konkurrenz z. B. zu CLARIN, DARIAH oder NEDIMAH, sondern ist ein wichtiger Schweizer Beitrag zur internationalen Kooperation im Bereich der geisteswissenschaftlichen Forschungsinfrastrukturen. Das Projekt *histHub.net* spielt in diesem Umfeld eine entscheidende Rolle; indem es in verschiedenen namhaften Schweizer Forschungsprojekten vorhandene Daten zusammenführt, sie über standardisierte Schnittstellen zugänglich macht und eine klare Entwicklungsperspektive – auch für weitere Projekte – aufzeigt, werden Insellösungen überwunden und damit überhaupt erst die Voraussetzungen für den effektiven Einsatz nationaler und internationaler Forschungsinfrastrukturen geschaffen. Bei den europäischen Forschungsinfrastrukturprojekten fehlt eine vergleichbare Integration bislang, so dass die Schweiz eine wichtige Vorreiterrolle im Bereich der geisteswissenschaftlichen Forschungsinfrastrukturen übernehmen könnte.

[8] Vgl. http://labeling.i3mainz.hs-mainz.de/
[9] Vgl. https://www.youtube.com/channel/UCb1xb5HGmZ1QEcFYwC9O0hA

Kuration: Migration und Qualitätssicherung digitaler Forschungsdaten

Iris Vogel [iris.vogel@uni-hamburg.de], Universität Hamburg

Zusammenfassung
Dieser Beitrag widmet sich möglichen Methoden, Mitteln und Problemen in der Kuration von Daten und Anwendungen aus abgeschlossenen Forschungsprojekten. Anhand von Beispielen aus den derzeitigen Kurationsvorhaben des gwin-Projekts an der Universität Hamburg werden die Vorgehensweisen von der Aquise und Analyse bis hin zur eigentlichen Migration in die Forschungsdateninfrastruktur vorgestellt und die Frage diskutiert, was die Kuration von Forschungsdaten und -anwendungen leisten soll und kann.

Abstract
Der permanente Wandel in Datenformaten und Präsentationsstandards, aber auch personelle Veränderungen sowie fehlende infrastrukturelle Maßnahmen zur Langzeitarchivierung, stellen eine potenzielle Gefahr für den Fortbestand digitaler Ressourcen dar. Eine Sensibilisierung dafür, dass die langfristige Verfügbarkeit von Forschungsdaten und -anwendungen aktive Bemühungen erfordert, ist jedoch in vielen geisteswissenschaftlichen Disziplinen erst im Entstehen begriffen. Dadurch ist eine Generation von Forschungsdaten und -anwendungen an einem Punkt angelangt, an welchem ihr Fortbestand als nicht gesichert gilt.

Die Kuration von Forschungsdaten abgeschlossener Projekte ist ein Teil des Service-Angebotes des gwin-Projekts der Fakultät für Geisteswissenschaften an der Universität Hamburg. Eine erste Bestandsaufnahme an der Fakultät diente zunächst der Beschreibung vorhandener Forschungsdaten und -anwendungen, ihres Kontextes sowie ihrer Form und Funktion. Das Ergebnis zeigt eine heterogene Forschungsdatenlandschaft, in welcher Formate und Repräsentationen der Daten und Anwendungen meist nicht mehr den aktuellen Standards entsprechen.

Ziel der Kuration ist in den meisten Fällen nicht nur die Archivierung der Daten selbst, nach Möglichkeit soll auch die Verfügbarkeit von Daten und Funktionalität für einen weiteren Lebenszyklus gesichert werden. Die aus der Kuration gewonnenen Erfahrungen geben überdies wichtige Impulse für die anderen Projektaufgaben, insbesondere die Beratung für die Umsetzung von datengestützten Forschungsvorhaben sowie den Ausbau der Forschungsinfrastruktur. Erfolgt im Zuge dessen eine Migration in die im Aufbau befindliche Forschungsdateninfrastruktur des Projektes, ist in der Regel auch eine Überarbeitung von Datenstrukturen und Funktionalität Teil der Aufgabe.

Die akute Notwendigkeit zur Kuration von Forschungsdaten kann unterschiedlich begründet sein. Zu den häufigsten Motiven gehören technische Unzulänglichkeiten, ggfs. verbunden mit potenziellen Sicherheitsrisiken sowie personeller Wandel, beispielsweise bedingt durch den Wegfall der Verantwortlichen für eine Datensammlung oder Forschungsanwendung. Auch gestalterische Fragen können zum Tragen kommen, wenn etwa die Präsentation nicht mit den Vorgaben der Institution vereinbar ist.

Um die Daten in die zentrale Infrastruktur zu überführen, ist zunächst die Einwilligung der jeweiligen Verantwortlichen essentiell. Nicht selten ist hier eine gewisse Überzeugungsarbeit zu leisten. Angst vor einen Verlust der Kontrolle über die Daten, Qualitätseinbußen durch den Migrationsprozess sowie indirekten Kosten durch den Abzug von dezentralen Mittel in ein zentrales Projekt führen dazu, dass dem Migrationsangebot zum Erhalt der Forschungsdaten teilweise mit Skepsis begegnet wird.

Aufgabe der Kuration ist folglich, die Forschungsdaten, Funktionalität und Anforderungen zu prüfen und in Zusammenarbeit mit den Verantwortlichen einen Migrationsplan zu erarbeiten, der die Verfügbarkeit des wissenschaftlichen Mehrwerts der Daten und Applikationen für eine weitere Lebensperiode (i.d.R. 10 Jahre) erhält und gleichzeitig den Aufwand für die zukünftige Pflege minimiert. Da die Forschungsdaten- und Anwendungen die Heterogenität der Fächer und Projekte in den Geisteswissen-schaften widerspiegeln, die Forschungsdateninfrastruktur jedoch versucht die Anzahl der Plattformen und Formate auf ein sinnvolles Maß zu beschränken, sind Kompromisse unabdingbar.

In den einfachsten Fällen kann Kuration bedeuten, auf technischer Ebene die Verantwortung für die Forschungsdaten bzw. -anwendung zu übernehmen. Werden Forschungsdaten jedoch in ein anderes System beziehungsweise auf eine andere Plattform migriert, muss sichergestellt sein, dass die Datenintegrität und Funktionalität der Anwendung auf dem Zielsystem Bestand hat. Mit Hilfe von automatisierten Testmethoden können Anforderungen in einer Form spezifiziert werden, die zu messbaren Ergebnissen führt und in ihrem begrenzten Rahmen reproduzierbar ist. Auf diese Weise kann sichergestellt werden, dass die Datenmigration in das Zielsystem erfolgreich war. Im Idealfall können diese Methoden der Qualitätssicherung auch dazu dienen, bei notwendigen Anpassungen im Zielsystem eventuelle Probleme aufzuspüren.

Ein wesentlicher Nachteil automatisierter Verfahren der Qualitätssicherung ist, dass diese zunächst einen gewissen Overhead generieren. Die reine Implementation von Testverfahren bzw. eine hohe Testabdeckung (gemessen an den Zeilen Testcode in Relation zu Applikationscode) alleine ist kein Garant für eine mittel- oder langfristige Qualitätssicherung. Es ist die sorgfältige methodische Konzeption, die den eigentlichen Aufwand verursacht und gleichzeitig darüber entscheidet, wie nachhaltig die Methoden der Qualitätssicherung auch für die Langfristverfügbarkeit nutzbar sind. Die Überprüfung der Datenintegrität kann überdies aufzeigen, inwieweit die Quelldaten in einem konsistenten Zustand übergeben wurden. Hier stellt sich die Frage, wie man mit fehlerbehafteter, fehlender oder veralteter Information umgeht, insbesondere wenn die inhaltlichen Ansprechpartner nicht mehr zur Verfügung stehen. Geht es bei der Kuration um das reine „Einfrieren“ eines Zustandes? Oder ist sie als eine Weiterentwicklung zu verstehen, welche dazu berechtigt Änderungen vorzunehmen, die als Optimierungen im Sinne der Nachhaltigkeit zu verstehen sind? Anhand mehrerer Kurations-Szenarien wird die Frage diskutiert, was die Kuration von Forschungsdaten und -anwendungen leisten soll und kann.

GeSIG – ein Projekt im Spannungsfeld einer sich entwickelnden Infrastrukturlandschaft

Cordula Meißner [cordula.meissner@uni-leipzig.de], Universität Leipzig
Franziska Wallner [f.wallner@rz.uni-leipzig.de], Universität Leipzig

Zusammenfassung

Der Beitrag stellt ein geisteswissenschaftliches Forschungsprojekt vor, welches sich zum Ziel setzt, das gemeinsame sprachliche Inventar geisteswissenschaftlicher Disziplinen zu erfassen und in Vorbereitung auf seine lexikographische Aufbereitung zu systematisieren. Es wird aus der Projektperspektive der Umgang mit Forschungsdaten diskutiert und dabei insbesondere auf Fragen eingegangen, welche die nachhaltige Verfügbarkeit von Forschungsdaten und ihre Anbindung an eine Forschungsinfrastruktur bzw. an ein Datenzentrum sowie rechtliche Aspekte betreffen.

Abstract

Sprache ist in der Wissenschaft nicht nur ein Instrument, um Sachverhalte zu vermitteln, sondern spielt für das wissenschaftliche Denken eine konstitutive Rolle. Dies gilt insbesondere für den Bereich der Geisteswissenschaften, da hier selbst die Gegenstände der Forschung größtenteils sprachlich verfasst sind (vgl. Kretzenbacher 2010). Die nicht-terminologische, disziplinenübergreifend verwendete Wissenschaftssprache spiegelt dabei in besonderem Maße die in Sprache niedergelegten Erkenntnisprozesse wider und ist somit von wesentlicher wissenschaftsmethodologischer Bedeutung. Dieser Bereich wird unter dem Begriff der allgemeinen oder auch alltäglichen Wissenschaftssprache zusammengefasst (Schepping 1976, Ehlich 1999). Durch ihren disziplinenübergreifenden Charakter fungiert die allgemeine bzw. alltägliche Wissenschaftssprache zudem als Bindeglied zwischen den einzelnen geisteswissenschaftlichen Fachdisziplinen. Darüber hinaus bildet sie die Grundlage für den Transfer wissenschaftlicher Erkenntnisse in die Gesellschaft. Eine systematische Erschließung und Beschreibung dieses Sprachbereichs für die Geisteswissenschaften steht jedoch bislang noch aus. Dies ist jedoch ein dringendes Desiderat, bildet es doch eine wesentliche Grundlage für die Wissenschaftspropädeutik und damit eine entscheidende Voraussetzung für eine erfolgreiche Teilhabe am deutschsprachigen Wissenschaftsdiskurs.

Das Projekt GeSIG (Das gemeinsame sprachliche Inventar der Geisteswissenschaften) setzt sich daher zum Ziel, erstmals das Inventar der allgemeinen Wissenschaftssprache der Geisteswissenschaften auf empirischer Grundlage zu bestimmen und damit den Grundstein für ihre umfassende Erschließung zu legen. Hierzu werden aktuelle korpusmethodologische Werkzeuge und Erschließungsverfahren für den Bereich der Wissenschaftssprachforschung fruchtbar gemacht, um einen vollständigen systematischen Zugriff auf den allgemeinwissenschaftlichen Wortschatzbestand der Geisteswissenschaften zu ermöglichen. Das allgemein-wissenschaftssprachliche Inventar der Geisteswissenschaften soll u.a. nach Lemmabestand, Frequenz, Wortklassen, Wortbildungsmustern und Wortfeldern erschlossen werden.

Ein auf diese Weise bestimmtes Inventar stellt eine wertvolle Grundlage für die Dokumentation und Erforschung der Sprache der Geisteswissenschaften dar und bietet die Ausgangsbasis für die Reflexion spezifisch geisteswissenschaftlicher Erkenntnisprozesse. Das Projekt ist als Pilotprojekt angelegt und soll Vorarbeiten liefern für den Aufbau einer umfassenden lexikographischen Ressource dieses Sprachbereichs.

In Anbetracht des avisierten langfristigen Ziels stellen sich Fragen nach einer nachhaltigen Sicherung der Projektergebnisse in besonderer Weise. Zentral ist dabei die Frage nach der Anbindung an ein Datenzentrum, um eine dauerhafte Nachnutzbarkeit auch unabhängig von der Verfügbarkeit von Fördermitteln zu gewährleisten. Daneben ist bereits in der aktuellen Projektphase zu klären, welche der entstehenden Forschungsdaten für eine Nachnutzung geeignet sind. Dies beinhaltet auch die Klärung rechtlicher Fragen – vor allem im Hinblick auf die verwendeten Primärdaten (vgl. DFG 2015). Im Beitrag werden mögliche Wege im Umgang mit diesen Fragen skizziert und offene Punkte zur Diskussion gestellt.

Literatur

DFG (2015), *Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora*. Verfügbar unter http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_recht.pdf [02.04.2015].

Ehlich, Konrad (1995), „Die Lehre der deutschen Wissenschaftssprache: sprachliche Strukturen, didaktische Desiderate". In: Heinz Kretzenbacher / Harald Weinrich (Hgg.): *Linguistik der Wissenschaftssprache*. Berlin, New York: de Gruyter, 325-351.

Kretzenbacher, Heinz (2010), „Fach- und Wissenschaftssprachen in den Geistes- und Sozialwissenschaften". In: Hans-Jürgen Krumm / Christian Fandrych / Britta Hufeisen / Claudia Riemer (Hgg.), *Deutsch als Fremd- und Zweitsprache. Handbücher zur Sprach- und Kommunikationswissenschaft* (HSK) (Bd. 35.1). Berlin, New York: de Gruyter, 493-501.

Schepping, Heinz (1976), „Bemerkungen zur Didaktik der Fachsprache im Bereich des Deutschen als Fremdsprache". In: Dietrich Rall / Heinz Schepping / Walter Schleyer (Hgg.): *Didaktik der Fachsprache. Beiträge zu einer Arbeitstagung der RWTH Aachen vom 30.9. bis 4.10.1974.* Bonn-Bad Godesberg: DAAD, 13-34.

Poster und Systempräsentationen

Big data in den „sozialen Medien“

Qualitative Forschungsstrategien und ihre informationstechnischen Herausforderungen am Beispiel von Twitter

Roman Knipping-Sorokin, roman.knipping-sorokin@uni-hamburg.de, Universität Hamburg
Gertraud Koch, gertraud.koch@uni-hamburg.de, Universität Hamburg
Samantha Lutz, samantha.lutz@uni-hamburg.de , Universität Hamburg

Zusammenfassung

Der Beitrag arbeitet am Beispiel von Twitter die informationstechnischen Fragen und Probleme heraus, die sich im Zusammenhang mit qualitativen Datenanalysen in den „sozialen Medien“ stellen.

Abstract

Forschung in den „sozialen Medien“ sind eng mit dem Stichwort "Big Data" und damit quantitativen Strategien der Auswertungen verbunden. Aber auch für die qualitative Sozial- und Kulturanalyse stellen die „sozialen Medien“ interessante Forschungszugänge dar. Es stellen sich hinsichtlich der hier vorzufindenden Forschungsdaten allerdings ebenfalls vielfältige Fragen und Herausforderungen. Viele dieser Fragen sind informationstechnischer Art, etwa wenn es um die Gewinnung und Speicherung der Daten zur weiteren Verarbeitung und Auswertung geht.

Am Beispiel von Twitter geht der Beitrag der Frage nach, wie die Daten aus den „sozialen Medien“ für qualitative Analysen gewonnen und aufbereitet werden können. Im Einzelnen wird es darum gehen, welche Art der Daten die qualitative Forschung dort vorfindet, wie diese selektiert werden können und welche Tools hierfür die Gewinnung der Daten zur Verfügung stehen. Dabei wird es zentral um die Möglichkeiten und Begrenzungen gehen, die die von Twitter angebotenen Schnittstellen mit sich bringen oder auch durch die Geschäftsmodelle des Unternehmens gesetzt werden und damit den möglichen Erkenntnishorizont für die qualitative Forschung abstecken. Es werden weiterhin die epistemologischen Implikationen der so gewonnen Daten reflektiert sowie ihre weitere Verwendung in der Kulturanalyse thematisiert. Auch hier zeigen sich eine Reihe informationstechnische wie auch epistemologische Fragen und Herausforderungen auf, wenn die Auswertung mittels QDA-Software erfolgt.

Forschungsdatenmanagement in interdisziplinären Forschungsverbünden

Thomas Burch [burch@uni-trier.de], Universität Trier. Trier Center for Digital Humanities
Gisela Minn [minn@uni-trier.de], Universität Trier. Forschungszentrum Europa
Marina Lemaire [marina.lemaire@uni-trier.de], Universität Trier. Servicezentrum eSciences
Yvonne Rommelfanger [rommelfanger@uni-trier.de], Universität Trier. Servicezentrum eSciences

Zusammenfassung

Forschungsdatenmanagement ist heute ein Teil der guten wissenschaftlichen Praxis. Professionelles Datenmanagement beginnt in der Planungsphase des Forschungsprojektes. Es werden die Ziele und Prinzipien des professionellen Forschungsdatenmanagements in einem interdisziplinären Forschungsverbund definiert, die Services und Leistungen, die Forscher benötigen, identifiziert und dargelegt, wie sich die einzelnen Lebensphasen des Data-Life-Cycle zum Forschungsprozess, der einzusetzenden IT-Infrastruktur sowie deren Betrieb verhalten und welche Aufgaben mit den jeweiligen Phasen verbunden sind.

Abstract

Die strukturierte Aufbereitung und nachhaltige Sicherung von Forschungsdaten sowie die Entwicklung und Bereitstellung virtueller Forschungsumgebungen (VFU), die die orts- und zeitunabhängige Zusammenarbeit in Forschergruppen ermöglichen, gehören zu den zentralen Aufgabenfeldern bei der strategischen Neuausrichtung der wissenschaftlichen Informationsinfrastrukturen für die Geistes- und Sozialwissenschaften. Ausgehend von der Frage, wie die in den zahlreichen Digitalisierungsvorhaben sowie in Forschungsprojekten zunehmend in großem Umfang entstandenen digitalen Daten über den Projektabschluss hinaus angesichts des raschen technologischen Wandels gesichert werden können, entstanden Machbarkeitsstudien und umfassende Empfehlungen zur Langzeitarchivierung und zur Lösung der damit verbundenen technischen, organisatorischen und rechtlichen Herausforderungen. Zunächst konzentrierten sich diese Aktivitäten auf die Datensicherung beim Abschluss von Forschungsprojekten. Inzwischen wird dafür plädiert, das Datenmanagement bereits in der Planungsphase von Forschungsprojekten zu berücksichtigen und dabei alle für eine adäquate langfristige Nachnutzung der Forschungsdaten notwendigen Schritte in den Blick zu nehmen. Die dazu erforderlichen Aufgaben werden in Datenlebenszyklus-Modellen systematisch beschrieben und entsprechende Richtlinien zum Umgang mit Forschungsdaten einschließlich Metadaten sowie zur Erstellung von Datenmanagementplänen dokumentiert.

Das Poster präsentiert, wie an der Universität Trier mit der besonderen Herausforderungen des Forschungsdatenmanagements in interdisziplinären Forschungsverbünden umgegangen wird. So müssen schon in der Planungsphase die Anforderungen mehrerer Teilprojekte und Arbeitsgruppen unterschiedlicher Fächer berücksichtigt und zusammengeführt werden, um die Zusammenarbeit innerhalb des Projektverbundes zu unterstützen. Für die einzelnen Forschungsvorhaben müssen Forschungsdatenmanagementkonzepte entwickelt werden, die in einem verbundübergreifenden Konzept zusammengeführt

werden und der Data Policy gerecht werden. Die Planungs-, Konzeptions- und Dokumentationsarbeiten beziehen sich dabei auf den gesamten Forschungsprozess, d.h. von der Erfassung der Forschungsprimärdaten, über ihre Analyse, Publikation bis hin zur langfristigen Sicherung und Bereitstellung für künftige Forschungsvorhaben. Ebenfalls zu berücksichtigen sind die mit dem Datenmanagement verbundenen organisatorischen, technischen und rechtlichen Aspekte in den verschiedenen Phasen des Datenlebenszyklus, wobei insbesondere die für die Nachnutzung notwendigen Informationen in einem Datenmanagementplan dokumentiert werden. Ausgehend von einer detaillierten Analyse der Fragestellungen und Arbeitsmethoden der jeweiligen Teiluntersuchungen müssen in Zusammenarbeit mit den Teilprojekten Richtlinien für die Bearbeitung der Forschungsprimärdaten in den verschiedenen Arbeitsphasen und für die systematische Erfassung der für die Beantwortung der jeweiligen Forschungsfragen relevanten Sachinformationen, Meta- und Kontextdaten entwickelt werden. Dabei sind die durch die Disziplinenvielfalt im Verbund begründete Heterogenität der Dokumente, das breite Spektrum von Arbeitsmethoden sowie die Unterschiede in Formen und Intensität digital basierter Arbeitsweisen zu beachten. Es werden deshalb unterschiedliche Datenmanagementstrategien entwickelt, die sowohl die individuelle Forschungsarbeit berücksichtigen als auch die Anforderungen, die an eine spätere langfristige Verfügbarkeit und Nachnutzbarkeit gestellt werden.

Die jeweiligen Modelle zur Beschreibung der Primärdaten, ihrer Erhebung, Anreicherung mit strukturellen, deskriptiven und administrativen Metadaten sowie ihre Bearbeitung über die verschiedenen Phasen des Forschungsprozesses hinweg bilden die Basis für die jeweilige, spezifisch auf die Teiluntersuchungen abgestimmte Konfiguration der IT-Infrastruktur. Darüber hinaus müssen einheitliche Regeln zur Erfassung und Erschließung der Primärquellen erstellt werden, um diese innerhalb des Verbundes für die gemeinsame Forschungsarbeit austauschen und über die Projektlaufzeit hinweg langfristig sichern zu können.

MyCoRe – Eine Software zum Aufbau von Repositorien

Wiebke Oeltjen [wiebke.oeltjen@uni-hamburg.de], Universität Hamburg

Zusammenfassung

Auf dem Poster wird die Open-Source-Software „MyCoRe“ vorgestellt, die als Basis verschiedener Repositorien dient. Die Funktionalität der Software wird kurz geschildert, und mögliche Einsatzgebiete werden skizziert. Des Weiteren werden Charakteristika der Software genannt.

Abstract

MyCoRe ist eine Open-Source-Software, die als Framework zum Erstellen von fachspezifischen und institutionellen Repositorien dient. Seit 2001 wird die Software von einer bundesweiten Community bereitgestellt, kontinuierlich weiter entwickelt und an neue Anforderungen angepasst. Der Name – gesprochen [my core] – weist darauf hin, dass es sich um einen Softwarekern handelt, der in eigenen, individuell angepassten Repositorien eingesetzt werden kann („My“ Content Repository). In den Repositorien können Forschungsdaten, Publikationen oder andere digitale Dokumente (wie z.B. Urkunden, Akten, Bücher, Handschriften, Kataloge, Zeitschriften, Zeitungen etc.), aber auch Bilder, Audiodateien und Videos erfasst werden.

Das MyCoRe-Framework stellt ein breites Spektrum an Funktionen bereit, die für den Betrieb von digitalen Archiven, Online-Zeitschriftenservern oder Repositorien benötigt werden. Dazu gehört das Erschließen, Recherchieren, Veröffentlichen und Präsentieren von Metadaten und digitalen Inhalten verschiedenster Art. Suchfunktionen ermöglichen Recherchen in Metadaten, Volltexten und XML-Strukturen. Bildbetrachter gewähren Zugang zu Digitalisaten und PDF-Viewer zeigen Dokumente. Klassifikationen können als Standard importiert oder frei editiert werden. Für Dokumente jeder Art kann freier Zugriff (Open Access) gewährt oder eine Zugriffskontrolle (ACL) eingerichtet werden. Datenexporte sind in den verschiedenen Formaten möglich (z.B. Bibliografien in den Formaten BibTeX, MODS, RIS, ISI, MS Word bibliography oder Dublin Core). Des Weiteren dienen Schnittstellen und Protokolle, wie das OAI Protocol for Metadata Harvesting, für den automati-sierten Datenaustausch und die Vernetzung externer Informationsressourcen. So kann z. B. die „Gemeinsame Normdatei“ (GND) zur Datenvernetzung genutzt werden (Linked Open Data). Das Archivieren großer Datenmengen, wie z.B. Forschungsdaten, ist ebenso möglich wie langfristiges Speichern.

Folgende Charakteristika zeichnen MyCoRe aus:

- MyCoRe ist **vielseitig** einsetzbar: Bundesweit werden MyCoRe-Anwendungen z.B. als fachspezifische Repositorien, Archive, Kataloge, Dokumentenserver, Zeitschriftenserver oder Bilddatenbanken betrieben.
- MyCoRe ist **anpassbar** an eigene Anforderungen und an individuelle Datenmodelle für das jeweils eigene Repositorium – „*My Content Repository*“.
- MyCoRe ist **nachhaltig** aufgrund des Einsatzes bewährter Software-Komponenten und der kontinuierlichen Weiterentwicklung der Software.

Über 65 MyCoRe-Anwendungen werden bundesweit an mehr als 20 Standorten betrieben. Eingesetzt wird MyCoRe von Universitätsbibliotheken und Universitätsrechenzentren, aber auch von Bundeseinrichtungen wie dem Statistischen Bundesamt und dem Bundesministerium für Ernährung und Landwirtschaft (BMELV).

Die neueste Entwicklung ist das „MODS Institutional Repository MIR“, das als Softwarepaket zur Verfügung gestellt wird und heruntergeladen, installiert und genutzt werden kann. Der Webanwendung liegt das Datenmodell „Metadata Object Description Schema“ (MODS) von der Library of Congress (LoC) zugrunde.

Weitere Informationen sind auf der MyCoRe-Homepage unter www.mycore.de nachzulesen.

Weblinks:

- MIR: http://www.mycore.de/mir
- MIR-Installationsanleitung: http://mycore.de/documentation/getting_started/mir.html
- MyCoRe: http://www.mycore.de

Literatur:

- Lützenkirchen, F: *MyCoRe – Ein Open-Source-System zum Aufbau digitaler Bibliotheken*. In: Datenbank Spektrum 2(4), November 2002, S. 23–27
- Oeltjen, W.: *Digitale Archive und virtuelle Bibliotheken auf der Basis von MyCoRe*. In: Robertson-von Trotha, C. Y.; Schneider, R. H. (Hrsg.): *Digitales Kulturerbe. Bewahrung und Zugänglichkeit in der wissenschaftlichen Praxis*. KIT Scientific Publishing, Karlsruhe, 2015, S. 173–182

Archivierung, Management und Präsentation von geisteswissenschaftlichen Forschungsdaten mit GAMS und Cirilo

Elisabeth Steiner [elisabeth.steiner@uni-graz.at], Universität Graz

Zusammenfassung

Das Geisteswissenschaftliche Asset Management System (GAMS) dient gleichzeitig als Langzeitarchivierungsinfrastruktur und (Re-)Präsentationsplattform für Forschungsdaten aus den verschiedensten geisteswissenschaftlichen Disziplinen. Dabei stehen nachhaltige Verfügbarkeit, verlässliche Archivierung und flexible Nutzung der Ressourcen im Vordergrund. Mit dem Cirilo Client ist es möglich, Objekte auf Basis eines Inhaltsmodells zu erstellen, zu verwalten, Inhalte regelbasiert zu extrahieren und Präsentationsformen zuzuweisen.

Abstract

Archivierung: GAMS

Das Geisteswissenschaftliche Asset Management System (GAMS)[1] dient gleichzeitig als Langzeitarchivierungsinfrastruktur und (Re-)Präsentationsplattform für Forschungsdaten aus den verschiedensten geisteswissenschaftlichen Disziplinen. GAMS basiert auf dem *open source* Repositorium FEDORA Commons[2] und bietet damit die Vorteile dieser Softwarelösung wie OAIS-Konformität[3], Trennung von Form und Inhalt und vordefinierte Interfaces wie OAI-PMH[4]. Ein zentrales Merkmal von FEDORA ist sein Objektmodell, das die Erstellung von Inhaltsmodellen als „Schablonen" für Objekte im Sinne einer Klassendefinition erlaubt. Durch die XML-basierte Content-Strategie und die Verwendung standardisierter (Meta-)Datenformate entsteht eine nachhaltig verfügbare Sammlung wiederverwertbarer Forschungsdaten aus den Geistes- und Kulturwissenschaften.

Im wissenschaftlichen Kontext ist die persistente Zitierbarkeit der Ressourcen besonders wichtig, alle Datenobjekte im System erhalten daher einen permanenten Identifikator (PID) auf Basis des Handle-Systems.[5] Damit können digitale Objekte ähnlich einer Printpublikation eindeutig zitiert werden.

Seit 2014 ist GAMS nach den Kriterien des *Data Seal of Approval*[6] zertifiziert. Das Siegel wird an digitale Repositorien verliehen, die auf Grund ihrer Bewertung nach 16 Grundregeln als vertrauenswürdig eingestuft wurden (*trusted digital repositories*). Besonderes Augenmerk liegt dabei auf der langfristigen Verfügbarkeit, verlässlichen Archivierung und der Wiederverwendbarkeit der enthaltenen Ressourcen.

[1] http://gams.uni-graz.at
[2] http://fedora-commons.org
[3] http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf
[4] http://www.openarchives.org/pmh/
[5] www.handle.net
[6] www.datasealofapproval.org

Management und Präsentation: Cirilo Client

Cirilo ist ein Client, der für *content preservation* und *data curation* in FEDORA-basierten Repositorien entwickelt wurde. Er bietet Funktionalitäten wie Objekterstellung und –verwaltung, Versionierung, Normalisierung und unterstützt die Verwendung von standardisierten Datenformaten. Cirilo erleichtert das Ausführen von Massenoperationen wie Ingest oder Ersetzungen, die mit dem Client sehr einfach auf eine große Anzahl von Datenobjekten angewendet werden können. Cirilo bietet ein Set von vordefinierten Inhaltsmodellen, die ohne weitere Anpassungen für Archivierungsworkflows von Standardformaten wie TEI oder METS/MODS verwendet werden können. Dabei erfüllt das Werkzeug auch Aufgaben wie die Auflösung von Ortsnamen oder Ontologiekonzepten gegen definierte Normdateien und die automatische Extraktion semantischer Relationen aus den Inhaltsdaten. Die semantische Anreicherung und die semi-automatische Verknüpfung mit Normdaten erhöht die Datenqualität der Ressourcen mit Hinblick auf die Wiederverwendbarkeit und Nachnutzbarkeit. Benutzerdefinierte Mappings erlauben eine einfache Transformation der wichtigsten Inhalte in RDF-Triples und unterstützen damit die Verfügbarkeit als *linked open data*. Verschiedene Disseminationsoptionen können über den Client zugewiesen und verwaltet werden, dazu gehören neben projektspezifischen Präsentationsformen (HTML, PDF) auch Optionen wie Google Maps, Voyant Tools oder der DFG-Viewer.

Der Cirilo Client ist ein österreichischer Beitrag zum ESFRI-Projekt DARIAH[7] und steht damit einer breiten Öffentlichkeit zur Nachnutzung zur Verfügung.[8]

[7] www.dariah.eu
[8] Der Quellcode des Clients ist unter https://github.com/acdh/cirilo verfügbar, die Dokumentation und ein Tutorial kann unter http://gams.uni-graz.at/doku abgerufen werden. Ein „archive-in-a-box"-Installationspackage des Frameworks (Client inklusive einer FEDORA 3.5 Instanz) für Debian-basierte Linux-Distributionen befindet sich unter http://gams.uni-graz.at/download/cirilo-installer-2.4.tar.gz.

TraCES – eine digitale Forschungsinfrastruktur für die altätiopische Literatur

Cristina Vertan [cristina.vertan@uni-hamburg.de], Universität Hamburg

Zusammenfassung

In diesem Beitrag werden wir die Systemarchitektur sowie die Modelle für ein Wurzel- und ein Lemma-Lexikon für die Geʿez-Sprache erklären. Die Wurzel- und Lemma-Akquisition werden weitgehend durch computergestützte Prozesse realisiert.

Die erstellte Software wird bei der Präsentation des Beitrags vorgeführt.

Abstract

Dokumente in altäthiopischer Sprache (Geʿez) sind 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 einer Forschungsinfrastruktur für das Geʿez umso dringender. Das Ziel des Projekts TraCES ist primär die Entwicklung eines eRepositoriums: ein digitales Korpus der Geʿez-Sprache, zusammen mit Annotationen auf morphologischer, syntaktischer und semantischer Ebene. Dieses annotierte Korpus soll einerseits eine diachrone Analyse des Altäthiopischen ermöglichen, anderseits 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 Sammlung 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 Geʿez gewidmet.

Die Besonderheiten des südsemitischen Geʿez, u.a. die Austauschbarkeit der Laryngale 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 wohlgemerkt diese graphischen 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 entschieden und unternehmen somit unserer Kenntnis nach den ersten Versuch, eine semitische Sprache mit dem Lemon-Modell zu beschreiben.

Die Grundkomponenten eines Lemon-Lexikon-Modells für Geʿez 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.

In diesem Beitrag werden wir die Systemarchitektur sowie die Modelle für ein Wurzel- und ein Lemma-Lexikon für die Geʿez-Sprache erklären. Die Wurzel- und Lemma-Akquisition werden weitgehend durch computergestützte Prozesse realisiert.

Die erstellte Software wird bei der Präsentation des Beitrags vorgeführt.

Das TraCES Team: Alessandro Bausi (Projektleiter), Wolfgang Dickut (Philologie, Korpus-Annotation), Daria Elagina (OCR, Korpus-Annotation), Andreas Ellwardt (Morphologie und Lexikographie), Susanne Hummel (Philologie, Korpus-Annotation), Vitagrazia Pisani (Philologie, Korpus-Annotation), Eugenia Sokolinski (Projektkoordination, Webmaster), Cristina Vertan (Digitale Geisteswissenschaften, Computerlinguistik)

gwin – Aufbau einer fakultären Forschungsdaten-Infrastruktur

Kai Wörner [kai.woerner@uni-hamburg.de], Universität Hamburg

Zusammenfassung

Das gwin-Projekt [https://www.gwiss.uni-hamburg.de/gwin] arbeitet daran, zusammen mit dem Regionalen Rechenzentrum der Universität Hamburg eine Infrastruktur aufzubauen, die die langfristige Verfügbarkeit von Datenbanken, Repositorien und auch technischen Einrichtungen (z.B. Laboren) in den Geisteswissenschaften sicherstellt.

Abstract

Für das Problem der Langzeitarchivierung von Forschungsdaten gibt es je nach Wissenschaftsdisziplin und Art der digitalen Objekte bereits mehr oder weniger praktikable Lösungsansätze. So ist die Sicherung reiner Messdaten in den Naturwissenschaften eher ein Problem der Menge als der Methode, zum Teil existieren auch Datenzentren auf lokaler und internationaler Ebene zur Sicherung und Bereitstellung von Mess- und numerischen Daten.

In den Geisteswissenschaften existieren diese Strukturen noch nicht oder erst im Ansatz. Dies hat unterschiedliche Ursachen, darunter besonders die Heterogenität der Daten, Methoden und Begrifflichkeiten. Anders als in den Naturwissenschaften sind in den Geisteswissenschaften die Schritte im Forschungsprozess meist nicht deutlich voneinander getrennt, so dass nicht einfach zwischen Rohdaten bzw. Primärdaten, Zwischendaten und Ergebnisdaten unterschieden werden kann – die Datenbasis kann durch Annotation und Verknüpfung im Forschungsprozess durchaus noch erweitert und geändert werden.

Wichtiger noch: Häufig sind die Daten selber nur in der für sie vorgesehenen Softwareumgebung wirklich nutzbar, da die Verknüpfungen und Annotationen die Interpretation der Daten überhaupt erst möglich machen bzw. ihren eigentlichen Wert ausmachen. Somit sind auch die Forschungs-, Annotations- und Präsentationsumgebungen wichtige Ressourcen, die langfristig erhalten werden müssen.

Da solche Umgebungen in der Regel im Rahmen von zeitlich befristeten (Drittmittel-) Projekten entstehen, ist deren Pflege nach Projektende in den meisten Fällen nicht gewährleistet. Weder lokale IT-Support-Abteilungen noch das universitäre Rechenzentrum haben für solche Aufgaben Ressourcen frei, geschweige denn die fachliche Kompetenz, die häufig zum Verständnis der Anwendungen notwendig ist. An der geisteswissenschaftlichen Fakultät der Universität XXX sind zurzeit mehr als 60 unterschiedliche Anwendungen in Betrieb – vom weltweit genutzten Online-Lexikon bis zur kleinen lokalen Datenbank.

Das gwin-Projekt arbeitet daran, zusammen mit dem Regionalen Rechenzentrum der Universität Hamburg eine Infrastruktur aufzubauen, die die langfristige Verfügbarkeit von Datenbanken, Repositorien und auch technischen Einrichtungen (z.B. Laboren) in den Geisteswissenschaften sicherstellt. Um die konkreten Bedarfe und den Bestand der existierenden Anwendungen besser einschätzen zu können, wurde zunächst eine persönliche Befragung aller Professoren der Fakultät durchgeführt. Dabei wurde ermittelt, welche Forschungsdaten und -Anwendungen an der Fakultät bereits existieren, auf welcher technischen Basis sie basieren, welche Anwendungen und Datenerhebungen in

Zukunft geplant sind und wie es um die langfristige Verfügbarkeit dieser Daten und Anwendungen bestellt ist.

Das gwin-Projekt wird aufgrund der gewonnenen Informationen mit Leitlinien, Best-Practice-Empfehlungen, einem Angebot an ausgewählten technischen Lösungen und umfangreicher Beratung versuchen, den kompletten Lebenszyklus eines Forschungsprojektes zu begleiten und die langfristige Verfügbarkeit der entstehenden Forschungsdaten sowie die dauerhafte technische Betreuung der Anwendungen sicherstellen.

Zurzeit wird daran gearbeitet, das Fundament dieser technischen Lösungen auszubauen und zu stabilisieren und dabei gleichzeitig Datenbestände und Anwendungen, deren Fortbestand gefährdet ist, mittels Datenkuration in einen Zustand zu bringen, in dem sie langfristig weiter- und nachgenutzt werden können. Einige der Probleme, mit denen gleichgelagerte Initiativen zu kämpfen haben (insbesondere die Frage, wie die Lauffähigkeit der vielen, technisch heterogenen Anwendungen sichergestellt werden kann), sind auch in diesem Projekt nicht abschließend gelöst und können bei der Posterpräsentation diskutiert werden.

TeLeMaCo, ein kollaboratives Portal für Lehr- und Lernmaterialien

Hannah Kermes [h.kermes@mx.uni-saarland.de],
Jörg Knappen [j.knappen@mx.uni-saarland.de],
José Manuel Martínez Martínez [j.martinez@mx.uni-saarland.de],
Elke Teich [e.teich@mx.uni-saarland.de],
Mihaela Vela [m.vela@mx.uni-saarland.de],
Universität des Saarlandes

Zusammenfassung

Wir stellen TeLeMaCo, ein kollaboratives Portal für Lehr- und Lernmaterialien in den Sprachwissenschaften und Digital Humanities, vor. TeLeMaCo ist am CLARIN-D-Zentrum an der Universität des Saarlandes beheimatet und bietet einen strukturierten Zugriff auf einen großen Fundus an Lehr- und Lernmaterialien.

Das Portal ist leicht zu benutzen, sowohl für Besucher, die nach Lehr- und Lernmaterialien suchen, als auch für Nutzer, die ihre Materialien bekannt machen wollen.

Das Portal ist in die Infrastruktur des CLARIN-Projektes eingebunden.

Abstract

Einführung

Plattformen wie das Virtual Language Observatory (VLO, Van Uytvanck 2012) oder der CLARIN Federated Content Search Aggregator (`http://weblicht.sfs.uni-tuebingen.de/Aggregator/`) machen es möglich, in einer wachsenden Anzahl von Repositorien nach Sprachressourcen und Werkzeugen zur Sprachverarbeitung zu suchen. Für Lehr- und Lernmaterialien besteht diese Möglichkeit zur Zeit nicht. Ressourcen sind zumeist verstreut und nicht zentral zugreifbar. Ausnahmen gibt es derzeit nur für den Bereich der Dokumentation bedrohter Sprachen (EMELD School of best practices `http://emeld.org/school/index.html` und Resource Network for Language Documentation RNLD `http://www.rnld.org/resources`). TeLeMaCo (Teaching and Learning Materials Collection) soll helfen, diese Lücke zu schließen.

Der Dienst

TeLeMaCo ist ein Repositorium für unterschiedliche Arten von Lehr- und Lernmaterialien:

- Technische Dokumentation (Kurzanleitungen, Einführungen, Handbücher)
- Lernmaterialien zum Selbststudium (z.B. YouTube Videos oder Screencasts)
- Kurze Lehreinheiten (Dauer 2–4 Stunden), die in Vorlesungen integriert werden können
- Komplette Vorlesungen
- Referenzmaterialien (z.B. Fachwörterbücher)

Es ist möglich, auf im Netz verfügbare Materialien zu verlinken oder auch Materialien hochzuladen.

Materialien, die in TeLeMaCo eingestellt werden, werden mit Metadaten versehen. So ist es möglich, relevante Materialien schnell zu finden.

Einfache Benutzung

Die einfache Benutzbarkeit stand bei der Entwicklung von TeLeMaCo im Vordergrund.

So gibt es eine einfache und eine erweiterte Suche und die Möglichkeit, das Portal anhand von Schlagwörtern und Autoren zu durchsuchen.

Auch die Einstellung von neuem Lehr- und Lernmaterial ist einfach gehalten und dauert nicht mehr als fünf Minuten. Als Grundlage für die Beschreibung dient das Dublin-Core-Schema, das um einige Felder, die für Lehr- und Lernmaterialien relevant sind, erweitert wurde (z.B. Voraussetzungen, Zielgruppe, Lernziel, Niveau). Metadaten wie Schlagwörter und Autoren werden als Hyperlinks realisiert, dadurch entsteht eine Vernetzung unter den Lehr- und Lernmaterialien.

Sichtbarkeit

TeleMaCo wird von den großen Suchmaschinen (Google, MSN, Yahoo, Yandex, Baidu) indiziert. Die Sichtbarkeit der in TeLeMaCo enthaltenen Lehr- und Lernmaterialien wird dadurch noch zusätzlich erhöht. Auf der einen Seite werden Benutzer der Suchmaschinen über TeLeMaCo an ihr Ziel geführt, andererseits erhöht sich auch der Page Rank der Lehr- und Lernmaterialien selbst.

Evaluation

Seit September 2013 (Amoia et al. 2013) ist TeLeMaCo online. In der Zwischenzeit wurden mehr als 130 Materialien in den Sprachen Englisch, Deutsch und Französisch hinzugefügt. Für einige Werkzeuge (z.B. EXMARaLDA, Schmidt und Wörner 2008) ist die Abdeckung der verfügbaren Lehr- und Lernmaterialien schon sehr gut.

Zur Zeit kommen die meisten Materialien aus dem Umfeld von CLARIN-D, aber auch externe Benutzer werden auf TeLeMaCo aufmerksam.

Literatur

Marilisa Amoia, Hannah Kermes, Jörg Knappen, José Manuel Martínez Martínez, Elke Teich, and Mihaela Vela. 2013. TeLeMaCo—a collaborative repository for training and teaching materials in linguistics. In Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2013), Darmstadt, Germany.

Thomas Schmidt and Kai Wörner. 2009. EXMARaLDA – Creating, analyzing and sharing spoken language corpora for pragmatic research. Pragmatics, 19(4):565–582.

Dieter Van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the virtual language observatory. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

LAUDATIO

Ein flexibles Repositorium für historische Textdaten

Thomas Krause [krauseto@hu-berlin.de],
Anke Lüdeling [anke.luedeling@hu-berlin.de],
Carolin Odebrecht [carolin.odebrecht@hu-berlin.de],
Dennis Zielke [zielkede@cms.hu-berlin.de],
Humboldt-Universität zu Berlin

Keywords: historische Texte, Korpuslinguistik, (Wieder-)Nutzung von Forschungsdaten, Sichtbarkeit, Repositorium

Abstract

LAUDATIO ist ein frei zugängliches Repositorium zur persistenten Speicherung dieser historischen Texte und ihrer Annotationen. Dabei werden die importierten Korpora in ihrer ursprünglichen Struktur in einer permanenten und unveränderlichen Version gespeichert. In der Systemvorführung im Rahmen der Postersession werden die Funktionalitäten des vom Projekt LAUDATIO (weiter)entwickelten Repositoriums vorgestellt.

In den verschiedenen Geisteswissenschaften wie der Geschichtswissenschaft, der Literaturwissenschaft und der Linguistik stellen historische Texte wie Zeitungen, Briefe, Urkunden oder Predigten oft die zentrale Forschungsgrundlage dar, um Sprachstufen, geschichtswissenschaftliche Kontexte oder gesellschaftliche Gefüge einer bestimmten Epoche zu untersuchen.

Die aus der Forschungsarbeit entstandenen Korpora sollen neuen Projekten zugänglich gemacht und nachvollziehbar dokumentiert werden, um eine Wiederverwendung der Korpora von einer Disziplin in eine andere zu ermöglichen. So können Korpora, die beispielsweise für geschichtswissenschaftliche Projekte erhoben wurden, unter anderen Fragestellungen auch für linguistische Forschung interessant sein. Korpusprojekte aus verschiedenen Disziplinen stellen schon jetzt ihre Korpora mit insgesamt 2000 Texten und mit etwa zwei Millionen Wortformen über LAUDATIO mit freien Lizenzen zur Verfügung. Dazu wurde ein umfangreiches Forschungsdatenmodell und Metadatenschema (Odebrecht 2014) basierend auf TEI (Burnard & Rahtz 2004) entwickelt. Sofern die Korpora in Formaten vorliegen, die mit den Konvertierungstool Pepper (Zipser & Romary 2010) konvertiert werden können, werden sie auch auf unserer öffentlichen Installation der generischen Korpussuchmaschine ANNIS (Krause & Zeldes 2014) bereitgestellt. Durch die Kooperation mit diesen Digital Humanities Projekten stellen wir nicht nur die langfristige Speicherung der Daten dritter Projekte sondern auch deren Analyse sicher.

Um die fachübergreifende Nutzung der Korpora zu ermöglichen bestehen die Herausforderungen darin, dass Forschungsdatenmodell generisch für die verschiedenen Nutzer weiterzuentwickeln, sodass es als Grundlage für die erweiterten Funktionalitäten Indexierung, Facettierung und Anzeige der Suche, Import und Validierung und Anzeige der Korpora im LAUDATIO-Repositorium dient. Mit diesen Erweiterungen sollen die unterschiedlichen Eigenschaften der Korpora spezifisch und detailliert abgebildet werden.

Durch solche Anpassungen können bspw. Korpora verschiedener Sprachen wie das „Historische Syntax des Jiddischen“-Korpus oder ein koptisches Korpus des Projektes „Scriptorium“ in LAUDATIO integriert werden.

Unser Ziel ist es, im Rahmen der Initiativen der Digital Humanities verstärkt Korpora mit historischen Texten weiterer geisteswissenschaftlicher Disziplinen aufzunehmen und somit den Nutzerkreis von LAUDATIO zu erweitern.

Referenzen

Burnard, L., Rahtz, S. (2004) RelaxNG with Son of ODD. Extreme Markup Languages Proceedings 2004. Montréal.

Krause, Th., Zeldes, A. (2014) ANNIS3: A New Architecture for Generic Corpus Query and Visualization Literary and Linguistic Computing. http://dsh.oxfordjournals.org/cgi/content/abstract/fqu057?ijkey=GJBr0LhNfKW1g8i&keytype=ref

Odebrecht, C. (2014) Modeling Linguistic Research Data for a Repository for Historical Corpora. Digital Humanities Conference 2014. Lausanne. http://dharchive.org/paper/DH2014/Paper-737.xml

Zipser, F., Romary, L. (2010) A model oriented approach to the mapping of annotation formats using standards. Proceedings of the Workshop on Language Resource and Language Technology Standards 2010. Malta. http://hal.archives-ouvertes.fr/inria-00527799/en/

Fachspezifische Repositorien auf der Basis von MyCoRe

Wiebke Oeltjen [Wiebke.Oeltjen@uni-hamburg.de], Universität Hamburg
Hinnerk Rümenapf [Hinnerk.Ruemenapf@uni-hamburg.de], Universität Hamburg

Zusammenfassung

In diesem Papier wird eingangs die Open-Source-Software „MyCoRe“ kurz beschrieben. Im Weiteren werden exemplarisch drei Webanwendungen vorgestellt, die auf MyCoRe basieren. Diese fachspezifischen Repositorien werden in verschiedenen Bereichen eingesetzt. Sie unterscheiden sich in den Inhalten, im Datenmodell und Layout. Gemeinsam ist ihnen, dass sie Forschungsdaten enthalten. Die fachspezifischen Repositorien sollen in einer Systemvorführung gezeigt werden.

Abstract

Die Open-Source-Software MyCoRe wird seit mehr als 10 Jahren von einer kleinen Community an deutschen Universitätsbibliotheken und -Rechenzentren kontinuierlich weiterentwickelt. An mehr als 20 Standorten bundesweit werden über 65 Webanwendungen auf der Basis von MyCoRe betrieben. Der Name – gesprochen [my core] – weist darauf hin, dass es sich um einen Softwarekern handelt, der in eigenen, individuell angepassten Repositorien eingesetzt werden kann („My“ Content Repository). Der Softwarekern stellt Funktionen zum Erschließen, Recherchieren, Präsentieren und Archivieren der digitalen Forschungsdaten und Metadaten bereit. In den Repositorien können Forschungsdaten, Publikationen oder andere digitale Dokumente (wie z.B. Akten, Urkunden, Handschriften, Bücher, Kataloge, Zeitschriften, Zeitungen etc.), aber auch Bilddateien, Tondokumente und Videos erfasst werden.

Es werden exemplarisch drei fachspezifische Repositorien gezeigt, die Forschungsdaten enthalten. Sie geben einen Einblick in die Vielfalt der Repositorien, die auf der Basis von MyCoRe entwickelt werden:

1. Die Digitale Bibliothek „Bach Digital“ zeigt digitalisierte Handschriften, Handschriftenfragmente und andere Originalquellen der Werke Johann Sebastian Bachs. In einem umfangreichen Datenmodell werden Metadaten zu Bachs Werken, den Quellen, Schriftproben und Wasserzeichen erfasst. Quellen werden auf Werke, Schriftproben und Wasserzeichen verlinkt, die Editionen verweisen wiederum auf Werke. Mittels Klassifikationen (Normdateien) wird eine einheitliche Benennung von Sachverhalten sichergestellt. Schnittstellen ermöglichen einen Export in standardisierten Datenformaten, wie z.B. DublinCore, Marc21 XML oder MODS.
2. Der Online-Katalog „The International Treasury of Islamic Manuscripts“ (ITIM) enthält digitalisierte Manuskripte mit historischen orientalischen Handschriften, die als Digitalisate zur Verfügung stehen und mit einem Bildbetrachter online angesehen werden können. Dazu werden umfangreiche Metadaten in Arabisch und Englisch erfasst. Mittels Klassifikationen und Normdateien, wie z.B. den „*Library of Congress Name Authority Files*“ (LCNAF), werden die Manuskripte mit normierten Begriffen versehen. Neben der Katalogisierung der historischen Handschriften werden in dem System Transliterationen der arabischen Schrift von Titeln und Autorennamen erfasst.
3. Die Online-Bibliothek „OpenAgrar“, die sechs Bibliotheken der Ressortforschungseinrichtungen des Bundesministeriums für Ernährung, Landwirtschaft und Verbraucherschutz (BMELV-Forschung) vereint, enthält neben Publikationen auch Forschungsdaten.

Weblinks:

- „Bach Digital“: www.bach-digital.de
- MyCoRe-Homepage: www.mycore.de
- „OpenAgrar“: openagrar.bmel-forschung.de
- „The International Treasury of Islamic Manuscripts“: www.manuscript-treasury.net

Literatur:

[OpenAgrar] Carl-Ulrich Koehler, Thomas Storck, Benedikt Hummel: OpenAgrar, das Repositorium für Einrichtungen im Geschäftsbereich des BMELV. In: GBDL-Mitteilungen, Heft 58/2013, S. 15-24, https://openagrar.bmel-forschung.de/receive/openagrar_mods_00001946

[Bach Digital] Christine Blanken, Christiane Hausmann, Jens Kupferschmidt, Stefan Freitag et al.: Bach Digital. Digitalisierung der Werkautographen und der originalen Aufführungsmaterialien Johann Sebastian Bachs. Umsetzung des Projektes auf Basis der Content Management Anwendung des MyCoRe Arbeitskreises. Projektdokumentation, Version 2.7, Leipzig, 29/08/2013, http://www.bach-digital.de/BachDigital_Doku.pdf

[MyIHS] Thoralf Hanstein, Abde Ssammad Karmoun, Jens Kupferschmidt: MyIHS. Grundprojekt zur datenbankgestützten Erschließung und digitalen Bereitstellung orientalischer Handschriften Beschreibung der programmtechnischen Umsetzung basierend auf der Content Management Anwendung des MyCoRe Arbeitskreises. Version 2.11, Leipzig, 05.03.2015, http://www.manuscript-treasury.net/MyIHS_TechnicalGuide.pdf

FuD – Eine virtuelle Forschungsumgebung für die Geistes- und Sozialwissenschaften

Systemvorstellung

Thomas Burch [burch@uni-trier.de], Universität Trier. Trier Center for Digital Humanities
Gisela Minn [minn@uni-trier.de], Universität Trier. Forschungszentrum Europa
Marina Lemaire [marina.lemaire@uni-trier.de], Universität Trier. Servicezentrum eSciences
Yvonne Rommelfanger [rommelfanger@uni-trier.de], Universität Trier. Servicezentrum eSciences

Zusammenfassung

FuD ist eine virtuelle Forschungsumgebung für die Geistes- und Sozialwissenschaften. Sie unterstützt die zeit- und ortsunabhängige kollaborative Forschungsarbeit und ist vielseitig einsetzbar. Die modulare Software bildet den kompletten Forschungsprozess ab. Sie erleichtert den Datenaustausch mit anderen Forschern durch ein elaboriertes Rechtemanagement. FuD kann sowohl von großen Forschungsverbünden als auch von kleinen Forschungsprojekten eingesetzt werden.

Abstract

Das Softwaresystem FuD ist eine aus mehreren Teilkomponenten aufgebaute virtuelle Forschungsumgebung (VFU) mit umfangreichen Funktionalitäten. Es unterstützt die Zusammenarbeit in räumlich verteilten Arbeitsgruppen während der verschiedenen Phasen des Forschungsprozesses: von der Inventarisierung und Erfassung der Forschungsdaten, über deren Erschließung und Analyse, bis hin zur Ergebnispublikation und anschließender Datenarchivierung.

Die Inventarisierungskomponente bietet eine Vielzahl flexibel konfigurierbarer Eingabemasken zur Text- und Metadatenerfassung für unterschiedliche Dokumenttypen. Mechanismen zur Volltext- und Digitalisatsspeicherung, Dokumentorganisation und zum Datenaustausch stehen zur Verfügung ebenso wie Recherchetools sowie Import- und Exportschnittstellen. Ferner können externe Softwaretools angebunden werden, z. B. das Transkriptionswerkzeug Transcribo. So wurde mit FuD z. B. am Deutschen Historischen Institut in Paris die bislang noch unerschlossenen Archivbestände der Korrespondenz der Constance de Salm gesichtet und systematisch erfasst. Die Inventarisierungsdaten und die zugehörigen Digitalisate eines umfangreichen Teilkorpus können bereits online recherchiert werden.

In der Analysekomponente hält FuD eine Palette von Werkzeugen für die qualitative und quantitative Textanalyse bereit. Das Trierer Projekt „Sprachliche Konstruktion sozial- und wirtschaftspolitischer Krisen in der BRD von 1973 bis heute“ analysiert hiermit Argumentationsmuster und reichert ausgewählte Textpassagen mit Informationen zu Sprechhandlungen und Diskursakteuren an. Auf diese Weise wird der induktive Forschungsprozess unterstützt, da die Kategorienschemata fortwährend durch den Nutzer weiterentwickelt und umstrukturiert werden können. Auswertungsalgorithmen unter-

stützen die Zusammenführung der Teilergebnisse und den Prozess des Erkenntnisgewinns. Ferner können die Annotationstools für den Aufbau textkritischer Apparate genutzt werden, um so mit FuD Print- und Online-Editionen zu erstellen. So kann für die „Edition der Zunftordnungen für Maler bis um 1800“ der Trierer Arbeitsstelle zur Künstlersozialgeschichte der Rohsatz direkt aus FuD im LaTeX-Format exportiert werden. Ebenso können digitale Werkausgaben mit FuD über diverse Schnittstellen in unterschiedlichen Publikationssystemen erstellt werden, wie z. B. die digitale Edition der Korrespondenz August Wilhelm Schlegels.

Die Archivierungskomponente von FuD ermöglicht die Aufbereitung der Daten für die Langzeitarchivierung und den Export in entsprechende Repositorien. Hiermit können die in FuD erfassten und bearbeiteten Forschungsdaten längerfristig gesichert und bereitgestellt werden. Für die Universität Trier wurde eine Schnittstelle zum Virtuellen Datenrepositorium (ViDa) entwickelt.

Ziel des FuD-Konzeptes ist die Erstellung einer passenden eScience-Strategie abgestimmt auf die individuellen Bedürfnisse der Forschungsprojekte, die gemeinsam mit den beteiligten WissenschaftlerInnen entwickelt wird. Daher setzt sich das FuD-Team aus Fachwissenschaftlern und Informatikern zusammen, um zum einen bedarfsgerechte Softwarelösungen zu implementieren, zum anderen aber auch um die Forschenden bei der Planung und Durchführung von IT-gestützten Forschungsprojekten von der Antragstellung über die Integration der Software in den Forschungsprozess bis hin zur Aufbereitung der Daten für die langfristige Verfügbarkeit zu beraten.

FuD wird an der Universität Trier vom Servicezentrum eSciences für uni-interne wie auch externe Projekte angeboten, das auch weiterhin die koordinierte und nachhaltige Weiterentwicklung sowie den Regelbetrieb von FuD garantiert. Die für den SFB 600 „Fremdheit und Armut“ implementierte Arbeitsplattform FuD wurde im DFG-Projekt „FuD2015“, ein Kooperationsprojekt des Trier Center for Digital Humanities und des Forschungszentrum „Europa“, zu einer projektunabhängigen VFU weiterentwickelt und ein entsprechendes Finanzierungs- und Organisationsmodell entworfen, das nun vom eSciences-Zentrum der Universität Trier umgesetzt wird. Derzeit wird die Software in über 20 Projekten an Forschungsinstitutionen, Universitäten und Akademien in den unterschiedlichsten Anwendungsbereichen eingesetzt.

Datenbank für Gesprochenes Deutsch (DGD)

Thomas Schmidt [thomas.schmidt@ids-mannheim.de], Institut für Deutsche Sprache, Mannheim

Zusammenfassung

Die Datenbank für Gesprochenes Deutsch (DGD) ist die zentrale Plattform für den Zugriff auf Daten des Archivs für Gesprochenes Deutsch (AGD). Über die DGD werden über 20 mündliche Korpora des Deutschen im Gesamtumfang von mehr als 2.500 Stunden Audio für Online-Browsing und -Query angeboten.

Abstract

Die Datenbank für Gesprochenes Deutsch (DGD, http://dgd.ids-mannheim.de, Schmidt 2014a) ist die zentrale Plattform für den Zugriff auf Daten des Archivs für Gesprochenes Deutsch (AGD). Über die DGD werden über 20 mündliche Korpora des Deutschen im Gesamtumfang von mehr als 2.500 Stunden Audio angeboten.

Der Bestand umfasst erstens mehrere große Variationskorpora des Deutschen, insbesondere das Korpus „Deutsche Mundarten“ (Zwirner-Korpus) mit mehreren Satelliten-Korpora nach dem gleichen Design („Deutsche Mundarten: Ehemalige deutsche Ostgebiete“, „Deutsche Mundarten: DDR“ und mehrere kleinere, regional begrenzte Sammlungen von Dialektaufnahmen) sowie das Korpus „Deutsche Umgangssprachen“ (Pfeffer-Korpus). Diese Dokumentation binnendeutscher Mundarten wird komplementiert durch Korpora auslandsdeutscher Varietäten, z.B. das Korpus „Australiendeutsch“ und drei Korpora zum Emigrantendeutsch in Israel.

Zweitens bietet die DGD Zugriff auf verschiedene Gesprächskorpora, u.a. das Berliner Wendekorpus, die Korpora „Grundstrukturen“ (Freiburger Korpus) und „Dialogstrukturen“, sowie das Korpus „Elizitierte Konfliktgespräche“. Mit dem Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK, Schmidt 2014b) wird im AGD ein großes, breit stratifiziertes Gesprächskorpus des Deutschen aufgebaut, das technisch und methodisch auf dem Stand aktueller bester Praktiken ist und der wissenschaftlichen Gemeinschaft ebenfalls über die DGD zur Verfügung gestellt wird.

Die Bestände sind nach einem einheitlichen Metadatenschema dokumentiert und durch Transkriptions- und Annotationsdaten, die ebenfalls auf einem gemeinsamen Datenmodell basieren, für die wissenschaftliche Analyse erschlossen.

Die DGD erlaubt zum einen ein exploratives Browsen auf diesen Daten. Korpus-, Sprecher- und Ereignisdokumentationen können eingesehen und die zugehörigen Audiodateien online abgespielt werden. Mit dem Audio alignierte Transkripte werden dem Benutzer in einer HTML5-basierten Darstellung präsentiert, die das Anspringen beliebiger Stellen im Transkript und das synchronisierte Abspielen der entsprechenden Stelle der Aufnahme ermöglicht. Diese Form des Zugangs dient sowohl dem Kennenlernen der Datenbestände als auch dem Einstieg in deren qualitative Analyse.

Für die gezielte Auswertung der Daten in quantitativer Hinsicht bietet die DGD zum anderen mehrere Recherchefunktionen. Über eine strukturierte Metadatensuche können nach flexibel spezifizierbaren Kriterien (z.B.: Gespräche mit Sprechern aus dem norddeutschen Raum, älter als 40 Jahre) Teilmengen des Gesamtbestands ausgewählt und als virtuelle Korpora gespeichert werden. Die strukturierte Tokensuche erlaubt korpuslinguistische Anfragen über mehrere Annotationsebenen (Transkription,

orthographische Normalisierung, Lemmatisierung, POS-Annotation), deren Ergebnisse in vielfältiger Hinsicht kontextualisiert (d.h. mit Metadaten korreliert, auf Transkript- und Aufnahmekontext rückbezogen) werden können.

Bei allen Funktionen zum Browsen und Durchsuchen der Daten legt die DGD Wert darauf, korpusgesteuerte Analysemethoden zu ermöglichen, in denen Hypothesen aus den Daten selbst generiert und in einer interaktiven Auseinandersetzung mit selbigen schrittweise verfeinert werden können.

Die DGD ist seit Ende 2012 online und hat mittlerweile mehr knapp 4000 registrierte Nutzer aus Forschung und Lehre. Datenbestände und Funktionalität werden kontinuierlich erweitert. Ein Transfer der in der DGD erprobten Methoden auf andere Forschungsdatenzentren, die mit mündlichen Daten arbeiten, ist grundsätzlich denkbar.

Referenzen

Schmidt, Thomas (2014a): The Database for Spoken German – DGD2. In: Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC'14), Reykjavik, Iceland: European Language Resources Association (ELRA).

Schmidt, Thomas (2014b): The Research and Teaching Corpus of Spoken German – FOLK. In: Proceedings of the Ninth con-ference on International Language Resources and Evaluation (LREC'14), Reykjavik, Iceland: European Language Resources Association (ELRA).

Vorträge

- **Charakteristika sprachtypologischer Daten und ihre projektübergreifende Integration mit Hilfe von digitalen Ontologien** 1
  Alexander Borkowski, Andrea C. Schalley, Griffith University
- **Prämissen und Schlussfolgerungen der Angebotsentwicklung des Humanities Data Centres mit Hilfe des Konzepts des „Forschungsdatentyps“** 3
  Andreas Aschenbrenner, Stefan Buddenbohm, Sven Bingert, Claudia Engelhardt, Ulrike Wuttke, Georg-August-Universität Göttingen
- **Objekterfassung, Präsentationsportal, Forschungsumgebung – die Bereitstellung von musealem Content für Forschung und Wissenschaft** 6
  Dr. Michael Merkel, Frank Dührkohp, Verbundzentrale des GBV
- **Bereitstellung und Nachnutzung von Forschungsdaten in den Geisteswissenschaften** 7
  Benedikt Fecher, Deutsches Institut für Wirtschaftsforschung
- **CLLD - Dateninfrastruktur für Diversity Linguistics** 10
  Robert Forkel, Max-Planck-Institut für evolutionäre Anthropologie
- **Knora: A Generic Storage System and API for Accessing and Reusing Digital Qualitative Research Data in the Humanities** 13
  Benjamin Geer, Tobias Schweizer, Lukas Rosenthaler, Ivan Subotic, University of Basel
- **Vom digitalen Hausmeister zum international vernetzten Forschungsdatenzentrum. Die Entwicklung der digitalen Infrastruktur des Hamburger Zentrums für Sprachkorpora.** 15
  Daniel Jettka, Hanna Hedeland, Timm Lehmberg, Universität Hamburg
- **Empirische Daten für die Kunstgeschichte erzeugen mit Artigo Social Image Tagging** 18
  Harald Klinke, LMU München
- **Database of historical places, persons, keywords, and lemmas** 19
  Natalia Korchagina, Rechtsquellenstiftung des Schweizerischen Juristenvereins
- **INEL – ein Infrastrukturvorhaben zur Dokumentation indigener nordeurasischer Sprachen** 21
  Beáta Wagner-Nagy, Hanna Hedeland, Timm Lehmberg, Michael Rießler, HZSK
- **Wissenstransfer und Wissensressourcen: Support und Helpdesk in den Digital Humanities** 25
  Timm Lehmberg, HZSK
- **Die Entwicklung eines Geschäftsmodells für den Betrieb der Virtuellen Forschungsumgebung FuD. Ein Erfahrungs- und Praxisbericht** 28
  Marina Lemaire, Gisela Minn, Yvonne Rommelfanger, Thomas Burch, Universität Trier
- **Forschungsdaten in historischen Projekten oder die Quelle im Netz** 31
  Daniel Burckhardt, Anna Menny, Institut für die Geschichte der deutschen Juden
- **Relationale Datenbank und Webtechnologie im Dienst der Geolinguistik: eine Projektvorstellung** 33
  Susanne Oberholzer, Markus Kunzmann, Ludwig-Maximilians-Universität München
- **IANUS als fachspezifisches Forschungsdatenzentrum für die Altertumswissenschaften in Deutschland** 35
  Felix Schäfer, Maurice Heinrich, Deutsches Archäologisches Institut
- **Überlegungen zu einem Betriebsmodell für etablierte wissenschaftliche Software** 37
  Thomas Schmidt, Kai Wörner, Institut für Deutsche Sprache

- **Vom analogen Archiv zum digitalen Forschungsdatenzentrum** **40**
  Thomas Schmidt, Institut für Deutsche Sprache
- **histHub.net – Ein mehrsprachiger Metadatenhub für Linked Open Data zu historischen Entitäten aller Epochen und Regionen der Schweiz** **42**
  Pascale Sutter, Rechtsquellenstiftung des Schweizerischen Juristenvereins
- **Kuration: Migration und Qualitätssicherung digitaler Forschungsdaten** **44**
  Iris Vogel, Universität Hamburg
- **GeSIG – ein Projekt im Spannungsfeld einer sich entwickelnden Infrastrukturlandschaft** **46**
  Cordula Meißner, Franziska Wallner, Universität Leipzig

Poster

- **Big data in den sozialen Medien. Qualitative Forschungsstrategien und ihre informationstechnischen Herausforderungen am Beispiel von Twitter** **49**
  Roman Knipping-Sorokin, Gertraud Koch, Samantha Lutz, Universität Hamburg
- **Forschungsdatenmanagement in interdisziplinären Forschungsverbünden** **50**
  Gisela Minn, Yvonne Rommelfanger, Thomas Burch, Marina Lemaire, Universität Trier
- **MyCoRe – eine Software zum Aufbau von Repositorien** **52**
  Wiebke Oeltjen, Universität Hamburg
- **Archivierung, Management und Präsentation von geisteswissenschaftlichen Forschungsdaten mit GAMS und Cirilo** **54**
  Elisabeth Steiner, Universität Graz
- **TraCES – eine digitale Forschungsinfrastruktur für die altäthiopische Literatur** **56**
  Alessandro Bausi, Andreas Ellwardt, Cristina Vertan, Universität Hamburg
- **Aufbau einer fakultären Forschungsdaten-Infrastruktur** **58**
  Kai Wörner, Universität Hamburg

Systempräsentationen

- **TeLeMaCo, ein kollaboratives Portal für Lehr- und Lernmaterialien** **60**
  Hannah Kermes, Jörg Knappen, José Manuel Martínez Martínez, Elke Teich, Mihaela Vela, Universität des Saarlandes
- **LAUDATIO: Ein flexibles Repositorium für historische Textdaten** **62**
  Thomas Krause, Anke Lüdeling, Carolin Odebrecht, Dennis Zielke, Humboldt-Universität zu Berlin
- **Fachspezifische Repositorien auf der Basis von MyCoRe** **64**
  Wiebke Oeltjen, Hinnerk Rümenapf, Universität Hamburg
- **FuD – Eine virtuelle Forschungsumgebung für die Geistes- und Sozialwissenschaften – Systemvorstellung** **66**
  Gisela Minn, Yvonne Rommelfanger, Thomas Burch, Marina Lemaire, Universität Trier
- **Datenbank für Gesprochenes Deutsch** **68**
  Thomas Schmidt, Institut für Deutsche Sprache