UNIVERSITÄT BERN

KONFERENZABSTRACTS

DHd2017 Bern
Digitale Nachhaltigkeit

13.–18. Februar 2017

DHd digital humanities im deutschsprachigen raum

Digital Humanities im deutschsprachigen Raum (DHd)

DHd 2017

Digitale Nachhaltigkeit

Konferenzabstracts

Universität Bern
13. bis 18. Februar 2017

Burgergemeinde Bern

Schweizerische Akademie der Geistes- und Sozialwissenschaften
Académie suisse des sciences humaines et sociales
Accademia svizzera di scienze umane e sociali
Academia svizra da scienzas umanas e socialas
Swiss Academy of Humanities and Social Sciences

FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

SWITCH

Die Abstracts wurden von den Autorinnen und Autoren in einem Template erstellt und mittels des von Marco Petris, Universität Hamburg, entwickelten DHConvalidators in eine TEI konforme XML-Datei konvertiert.

Koordination der Publikation: Prof. Dr. Michael Stolz
Korrektur der Auszeichnung der Bibliographie und Konvertierung TEI nach PDF: Reto Baumgartner
TEI to PDF scripts: Karin Dalziel
https://github.com/karindalziel/TEI-to-PDF
Bearbeitete Version von Aramís Concepción Durán
https://github.com/aramiscd/dhd2016-boa
Konferenz-Logo: Regina Wittwer (reGains | Atelier für Grafik und Illustration)
Umschlaggestaltung: Simone Hiltscher

online verfügbar: http://www.dhd2017.ch

4. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V.

Beachten Sie bitte die ergänzenden
Abbildungen auf S. 304–306 am Ende des Bandes.

Plenarvorträge

Workshops

Panels

Vorträge

Poster

Plenarvorträge

Digitale Nachhaltigkeit: Mittel und/oder Zweck?

Seele, Peter
peter.seele@usi.ch
Università della Svizzera italiana, Schweiz

Digitalisierung und Nachhaltigkeit stellen zwei der thematisch wichtigsten Themenkreise und Treiber sowohl des gesellschaftlichen Diskurses als auch der akademischen Forschung dar. Dies betrifft nicht nur die sogenannten ‚harten' Wissenschaften, in denen naturwissenschaftliche Messungen von Nachhaltigkeitsthemen wie Klimawandel, Kohlendioxid Emissionen oder Biodiversität Gegenstand der Forschung darstellen. Die beiden Themenkreise Digitalisierung und Nachhaltigkeit haben in den letzten Jahren auch die Kultur- und Geisteswissenschaften erreicht. Der Plenarvortrag geht auf diese Neuerung als Form der Kombination von Digitalisierung und Nachhaltigkeit in den Humanities und hier insbesondere in den Digital Humanities ein.

„Digitale Nachhaltigkeit" als emergentes Thema und Konzept lässt sich dabei in zwei Haupttypen unterteilen, so der Vorschlag dieser Keynote:

1. Digitale Nachhaltigkeit als *Mittel*. Dies bedeutet, dass Digitalisierung als Mittel verstanden wird, nachhaltige Entwicklung zu fördern. Wie lassen sich also Big Data und Co dazu einsetzen, Nachhaltigkeit zu fördern?

2. Digitale Nachhaltigkeit als *Zweck*: Dies bedeutet, dass das Digitale an sich in einer Weise zu gestalten wäre, die nachhaltig zu nennen wäre. In diesem Sinne wäre die Digitale Nachhaltigkeit der Zweck.

Analog dazu liesse sich die Digitale Nachhaltigkeit als Topos der Digital Humanities skizzieren, wobei die Digital Humanities ebenso in der Unterscheidung nach Mittel und Zweck dargestellt werden können.

Beide Hauptpositionen werden im Vortrag dargelegt und anhand von Beispielen und ersten positionsbestimmenden Forschungsbeiträgen diskutiert. Schliesslich verdient insbesondere die normative Grundierung des Nachhaltigkeitsdiskurses in den Kultur- und Geisteswissenschaften besondere Beachtung, da Nachhaltigkeit als prädeliberatives Konzept bereits normativ positioniert ist und dementsprechend wissenschaftlich zu reflektieren wäre.

Ein Bild sagt mehr als tausend Worte, aber sagen tausend Pixel mehr als ein Wort?

Süsstrunk, Sabine
sabine.susstrunk@epfl.ch
Digital Humanities Instituts (DHI) der École polytechnique fédérale de Lausanne, Schweiz

In diesem Vortrag werde ich das Wort „Digital“ in Digital Humanities genauer erläutern. Was genau ist eigentlich „digital“? Aus der Sicht der Informatik kann „digital“ Information sein, die in einem Format kodiert ist, das für eine Berechnung geeignet ist. Aber ist diese Kodierung für die Geisteswissenschaften überhaupt geeignet? Die ASCII-Kodierung eines Wortes hat sich als sinnvoll erwiesen und wird somit ausgenutzt. Aber wie ist es mit den Pixeln, die eine zwei- oder dreidimensionale Szene kodieren und entweder ein altes Manuskript, eine Kinderzeichnung, die Interpretation der Klassik eines Kunsthistorikers oder ein berühmtes Jazzkonzert repräsentieren könnten?

Anhand von Beispielen aus der Forschung des Digital Humanities Instituts (DHI) der ETH Lausanne (EPFL) werde ich die Kodierung visueller Informationen diskutieren, den Reichtum der bildlichen Darstellung für die Geisteswissenschaften erläutern, aber auch über die noch zu bewältigenden Herausforderungen diskutieren, bis wir die visuelle Information so nutzen können wie das Wort.

Wenn Forschen ein nicht reproduzierbarer Prozess ist – Nachhaltigkeit als Herausforderung in der Archäologie

Fless, Friederike
praesidentin@dainst.de
Deutsches Archäologisches Institut, Deutschland

Ein Archäologe arbeitet sich bei einer Ausgrabung durch viele historische Schichten in die Tiefe. Dieser Prozess ist nicht umkehrbar, so dass der Dokumentation des Grabungsprozesses eine besondere Bedeutung zukommt. Wie aber sichert man solche Daten, die in vielfältigen Formaten heute digital erhoben werden, langfristig? Wie kann man diese Daten in einem geschlossenen Datenlebenszyklus für Nachnutzungen zur Verfügung stellen? In welcher Weise können wir mit der Vielfalt von Datenformaten umgehen? Diesen grundsätzlichen Fragen will der Vortrag ausgehend von einer konkreten Disziplin, der Archäologie, nachgehen und dabei auch die aktuellen Entwicklungen im Bereich des Forschungsdatenmanagements aufzeigen. Aktuelle Vorschläge, wie sie der Rat für Informationsinfrastruktur in Deutschland für die Entwicklung einer Nationalen Forschungsdateninfrastruktur publiziert hat, sollen dabei ebenso beleuchtet werden wie die dahinter stehende Geschichte von Informationsinfrastrukturen, auf der diese Vorschläge aufbauen. Um jenseits der grundlegenden Entwicklungen des Forschungsdatenmanagements und der Informationsinfrastrukturen auch konkrete Beispiele und Lösungsansätze für die Frage von Nachhaltigkeit zur Diskussion zu stellen, sollen die technischen Lösungen, die im Rahmen der digitalen Angebote des Deutschen Archäologischen Instituts, aber auch des DFG-Projektes IANUS (Forschungsdatenzentrum für die Langzeitsicherung archäologischer und altertumswissenschaftlicher Daten) vorgestellt werden.

Workshops

Annotieren und Publizieren mit DARIAH-DE und TextGrid

Kollatz, Thomas
kol@steinheim-institut.org
Steinheim-Institut für deutsch-jüdische Geschichte Essen, Deutschland

Hegel, Philipp
hegel@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Veentjer, Ubbo
veentjer@sub.uni-goettingen.de
Niedersächsische Staat- und Universitätsbibliothek Göttingen, Deutschland

Söring, Sibylle
sibylle.soering@sub.uni-goettingen.de
Niedersächsische Staat- und Universitätsbibliothek Göttingen, Deutschland

Funk, Stefan E.
funk@sub.uni-goettingen.de
Niedersächsische Staat- und Universitätsbibliothek Göttingen, Deutschland

Annotieren und Publizieren mit DARIAH-DE und TextGrid

Im Rahmen des halbtägigen Workshops werden den Teilnehmerinnen und Teilnehmern Werkzeuge zum Publizieren und Annotieren von Forschungsdaten demonstriert, die im Rahmen von Hands-On-Einheiten anhand eigener und / oder bereitgestellter Daten erprobt werden können.

Vorgestellt und angewendet werden das TextGrid- und DARIAH-DE Repositorium, der DARIAH-DE Publikator und die DARIAH-DE Annotation Sandbox. Zudem wird in die Arbeit mit dem Text-Bild-Link-Editor des TextGrid Laboratoriums eingeführt und exemplarisch gezeigt, diese Text-Bild Relationen mit Hilfe des Web-Publikationstools „SADE – Scalable Architecture for Digital Editions“ in eine digitale Präsentation bzw. ein Web-Portal zu übernehmen.

Der Workshop richtet sich an Geisteswissenschaftlerinnen und –wissenschaftler aus text- und bildbasierten Disziplinen aller Phasen des akademischen Werdegangs ebenso wie an Vertreterinnen und Vertreter von Institutionen – etwa Bibliotheken, Forschungsverbünde oder Archive –, die im Rahmen ihrer Vorhaben digitale Forschungsinfrastruktur nutzen bzw. nutzen wollen, um ihre Forschungsdaten nachhaltig digital zu publizieren und zu annotieren.

Der Workshop liefert durch Kurzvorträge und Hands-On-Einheiten Einblicke in verschiedene Verfahren, Anwendungen und Workflows liefern, um Geisteswissenschaftlerinnen und Geisteswissenschaftlern die maschinenlesbare Annotation von Text- und Bilddaten sowie die Publikation solcher Forschungsdaten in einem Repositorium zu ermöglichen. Nach einer kursorischen Einführung in die Angebote von TextGrid und DARIAH-DE liefert ein Überblick über das Annotieren in den digitalen Geisteswissenschaften verschiedene Anwendungsszenarien, -anforderungen, -modelle und -technologien. Dabei werden neben bereits bestehenden Angeboten wie dem TextGrid Text-Bild-Link-Editor auch neuere Entwicklungen wie die Annotation Sandbox und das DARIAH-DE Repositorium und seine Publish GUI (Publikator) demonstriert und in interaktiven Übungen durch die Teilnehmenden anhand eigener bzw. zur Verfügung gestellter Daten erprobt.

Der Workshop ist Teil zweier konzeptionell eigenständiger Einreichungen zu den Angeboten der digitalen Forschungsinfrastrukturen TextGrid und DARIAH-DE. [1] . Der Besuch beider Workshops ermöglicht eine grundlegende und umfassende Einführung in und Anwendung von Architektur, Tools, Diensten und Workflows zum Annotieren, Sammeln, Modellieren, Recherchieren und Publizieren geisteswissenschaftlicher Forschungsdaten.

Annotationen in den digitalen Geisteswissenschaften

Digitales Annotieren ist zentrale Praxis bei der Wissensgenerierung und variiert je nach spezifischer wissenschaftlicher Zielsetzung und Forschungsgegenstand. Verfahren des fachwissenschaftlichen digitalen Annotierens bilden heute eine der Kernanwendungen der Digital Humanities. Im Zentrum steht dabei ein weites Spektrum von Daten und / oder

Objekten, z.B. Texte, Bilder und Musik (Töne, Noten). Digitale Annotationen unterscheiden sich daher in Form, Funktion und Tragweite. Einführend werden die technischen Ebenen und theoretischen Dimensionen der digitalen Annotation in den Geisteswissenschaften exemplarisch erörtert. Die vermittelten Grundlagen können danach im Workshop praktisch angewandt werden.

Annotieren im Rahmen einer digitalen Infrastruktur

Forschungsinfrastrukturen wie TextGrid und DARIAH-DE haben zum Ziel, methodologische Fähigkeiten auf diesem Gebiet zu vermitteln, entsprechende Verfahren zu evaluieren bzw. bereitzustellen und die nachhaltige Anwendung dieser Verfahren in den Fachwissenschaften zu ermöglichen.

Die DARIAH-DE Annotation Sandbox (Beta) ermöglicht heute die Text- und Bildannotation der Bestände des TextGrid Repository. Darüber hinaus können beliebige Webseiten uber den DARIAH-DE Annotationsdienst annotiert werden. Zudem lässt sich der DARIAH-DE Annotationsdienst in eigene Webseiten einbetten; hierzu wurden die digitalen Werkzeuge Annotator.js, Via und ein Annotation Manager über die DARIAH AAI (Authorization and Authentification Service) verfügbar gemacht.

Die DARIAH-DE Annotation Sandbox gestattet die direkte Verbindung der in den Repositorien publizierten Forschungsdaten mit ihrer digitalen Annotation. Diese schließt sowohl die disziplinübergreifenden Nachnutzung als auch die Datenanreicherung oder die Analyse ein. Mittelfristig können Annotationen somit als Zwischenschritt des Forschungsprozesses, aber auch als genuines Forschungsergebnis - etwa im Sinne einer Mikropublikation - verstanden bzw. generiert, verfügbar gemacht und als solches nachgenutzt werden. Im Rahmen einer digitalen Infrastruktur fließen sie wie die Forschungsdaten, auf die sie Bezug nehmen, ebenfalls in die Archivierung ein, um weiterverarbeitet und nachgenutzt zu werden.

Bilder in TextGrid annotieren

Ein weiteres Anwendungsszenario digitaler Annotation stellt die Annotation von Bildern bzw. Bilddaten dar. Eine Vielzahl von Werkzeugen im TextGrid Laboratory erlaubt das Arbeiten mit Texten und Bildern, aber auch beispielsweise mit Noten und Digitalisaten. Eine dieser Komponenten, die auch für die Annotation von Bildbereichen dienen kann, ist der Text-Bild-Link-Editor. Er unterstützt den in TextGrid integrierten XML-Editor bei der Alignierung von Text- und Bildelementen. Ziel ist die Erstellung einer Ausgabedatei, die die Textelemente und die topographische Position von rechtwinkligen und polygonen Bildbereichen in SVG miteinander verknüpft, wie dies zum Beispiel bei der Verbindung von Faksimiles und Transkriptionen in kritischen Editionen der Fall ist. Auch können Bilder auf diese Weise im Rahmen kunsthistorischer Untersuchungen annotiert werden.

Text-Bild-Relationen publizieren

Die Software SADE der Berlin-Brandenburgischen Akademie der Wissenschaften ist als „Skalierbare Architektur für digitale Editionen“ in TextGrid eingebunden, um eigene Webportale für die Publikation gestalten zu können. Sie enthält ein Modul, mit dem die Verknüpfungen, die mit dem Text-Bild-Link-Editor erstellt wurden, in ein Web-Portal übernommen werden können. Dieses Modul basiert auf dem in DARIAH-DE integrierten Werkzeug „Semantic Topological Notes” (SemToNotes). Es erlaubt unter anderem, Zeilen auf einem Digitalisat auszuwählen und Transkriptionen anzuzeigen.

Publizieren via Infrastruktur: Das DARIAH-DE Repositorium und der DARIAH-DE Publikator

Das DARIAH-DE Repositorium bildet eine zentrale Komponente der Infrastruktur, auf die mittels verschiedener Dienste und Anwendungen zugegriffen werden kann. Das Repositorium erlaubt es, Forschungsdaten zu speichern, diese mit Metadaten zu versehen und die Forschungsdaten durch die Generische Suche aufzufinden. Die Daten werden im DARIAH-DE Storage sicher gespeichert. Darüber hinaus ermöglicht das Repositorium die nachhaltige und sichere Archivierung von Datensammlungen bzw. Kollektionen.

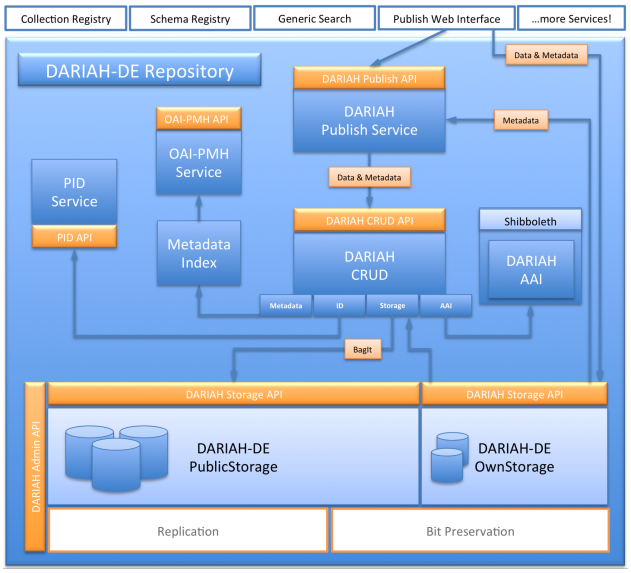


Abb.1: DARIAH-DE-Repositorium: Architektur

Dies ist komfortabel und intuitiv über ein Web-Interface des DARIAH-DE Portals im Browser möglich, dem DARIAH-DE Publikator. Daten im Repositorium sind in Kollektionen organisiert, die zunächst vom Nutzer über den Publikator angelegt und mit Metadaten ausgezeichnet werden. Einer Kollektion können beliebig viele Dateien zugeordnet werden, die ebenfalls über den Publikator hochgeladen und mit Metadaten ausgezeichnet werden. Eine publizierte Kollektion sowie alle darin enthaltene Objekte können unmittelbar nach dem Publizieren per Persistent Identifier (PID) referenziert werden und sind damit öffentlich zugänglich und nachhaltig referenzier- und zitierbar. Im nächsten Schritt kann die Kollektion in der Collection Registry nachgewiesen und veröffentlicht werden. Sobald die Kollektion selbst ebenfalls in der Collection Registry publiziert wurde, sind die Daten auch mit der Generischen Suche recherchierbar.

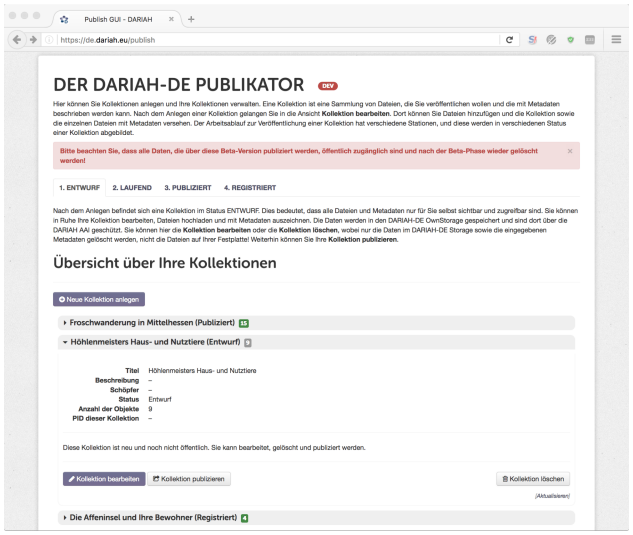


Abb. 2: DARIAH-DE Publikator: Übersicht über die Kollektionen

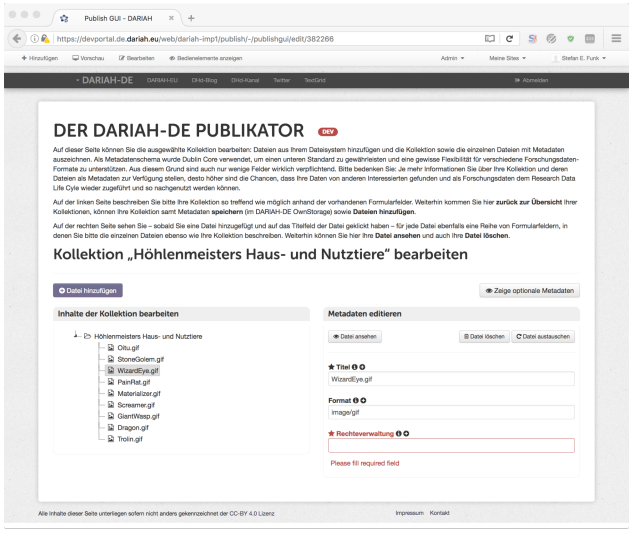


Abb. 3: DARIAH-DE Publikator: Kollektion bearbeiten

Anforderungen

Im Workshop werden exemplarisch Annotationen an einem Digitalisat in TextGrid vorgenommen. Zu diesem Zweck ist ein eigener Rechner mitzubringen, auf dem im Idealfall TextGrid bereits installiert ist – https://textgrid.de/download .

Eine Registrierung für TextGrid und DARIAH kann online beantragt werden unter

http://auth.dariah.eu/

Bitte teilen Sie uns im Vorfeld des Workshops (möglichst bis zum 5. Februar 2017) mit, ob und welche eigenen Materialien Sie verwenden wollen.

Für Rückfragen erreichen Sie uns unter workshop@de.dariah.eu

Kontaktdaten

Mirjam Blümm, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Abt. Forschung und Entwicklung, Papendiek 14, 37073 Göttingen, bluemm@sub.uni-goettingen.de
Forschungsinteressen: Virtuelle Forschungsumgebungen, Digitale Forschungsinfrastrukturen, Digitale Editionen

Stefan E. Funk, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Abt. Forschung und Entwicklung, Papendiek 14, 37073 Göttingen, funk@sub.uni-goettingen.de
Forschungsinteressen: Forschungsdatenmanagement, Digitale Langzeitarchivierung, Repositoriums-Technologien.

Canan Hastik, Technische Universität Darmstadt, Dolivostraße 15, Institut für Sprach- und Literaturwissenschaft, 64293 Darmstadt, hastik@linglit.tu-darmstadt.de
Forschungsinteressen: Digital Humanities, Semantisches Wissensmanagement, Digitale Kultur und Kunst

Philipp Hegel, Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Dolivostraße 15, 64293 Darmstadt, hegel@linglit.tu-darmstadt.de
Forschungsinteressen: Digitale Editionen, virtuelle Forschungsumgebungen

Thomas Kollatz, Salomon Ludwig Steinheim-Institut für deusch-jüdische Geschichte, Essen, Edmund-Körner-Platz 2, 42157 Essen, kol@steinheim-institut.org
Forschungsinteressen: Digitale Epigraphik, Jüdische Studien

Sibylle Söring, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Abt. Forschung und Entwicklung, Papendiek 14, 37073 Göttingen, soering@sub.uni-goettingen.de
Forschungsinteressen: Virtuelle Forschungsumgebungen, Digitale Forschungsinfrastrukturen, Digitale Editionen

Ubbo Veentjer, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Abt. Forschung und Entwicklung, Papendiek 14, 37073 Göttingen, veentjer@sub.uni-goettingen.de
Forschungsinteressen: Digitale Forschungsinfrastrukturen, Text- und Bild-Annotation, Visualisierungstechnologien.

Zahl der möglichen Teilnehmerinnen und Teilnehmer.

Aufgrund des hohen Praxisanteils soll die Zahl der Teilnehmerinnen und Teilnehmer auf möglichst 25 beschränkt bleiben.

Angaben zu einer etwa benötigten technischen Ausstattung.

WLAN / Beamer / Stellwände / Verlängerungskabel

Fußnoten

1. Siehe auch Workshop "Daten sammeln, modellieren und durchsuchen mit DARIAH-DE“

Bibliographie

Becker, Rainer / Bender, Michael / Borek, Luise / Hastik, Canan / Kollatz, Thomas / Lordick, Harald / Mache, Beata / Rapp, Andrea / Reiche, Ruth / Walkowski, Niels-Oliver (2016): „Digitale Annotationen in der geisteswissenschaftlichen Praxis“, in: *Bibliothek – Forschung und Praxis* 40 (2): 186–199 https://www.degruyter.com/view/j/bfup.2016.40.issue-2/bfp-2016-0042/bfp-2016-0042.xml?format=INT .

Bender, Michael / Borek, Luise / Kollatz, Thomas / Reiche, Ruth (2015): "Wissenschaftliche Annotationen: Formen – Funktionen – Anforderungen", in: *DHd-Blog* http://dhd-blog.org/?p=5388 .

Borek, Luise / Reiche, Ruth (2014): „Round Table ‚Annotation von digitalen Medien“ (Veranstaltungsbericht), in: *DHd-Blog* http://dhd-blog.org/?p=3831 .

Blümm, Mirjam / Funk, Stefan E. / Söring, Sibylle (2015): „Die Infrastruktur-Angebote von DARIAH-DE und TextGrid“, in: *Information. Wissenschaft & Praxis* 66 (5–6): 304–312.

Neuroth, Heike / Rapp, Andrea / Söring, Sibylle (2015): *TextGrid: Von der Community für die Community – Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften*. Göttingen http://www.univerlag.uni-goettingen.de/handle/3/Neuroth_TextGrid .

Schmunk, Stefan / Funk, Stefan (2015): „Das DARIAH-DE- und das TextGrid-Repositorium: Geistes- und kulturwissenschaftliche Forschungsdaten persistent und referenzierbar langzeitspeichern“, in: *Bibliothek Forschung und Praxis* 40 (2): 213–221 10.1515/bfp-2016-0020

Söring, Sibylle (2016): „Technische und infrastrukturelle Lösungen für digitale Editionen: DARIAH-DE und TextGrid“, in: *Bibliothek Forschung und Praxis* 40 (2): 207–212 10.1515/bfp-2016-0040 .

CUTE: CRETA Unshared Task zu Entitätenreferenzen

Reiter, Nils
nils.reiter@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Blessing, Andre
andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Echelmeyer, Nora
nora.echelmeyer@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Koch, Steffen
steffen.koch@vis.uni-stuttgart.de
Universität Stuttgart, Deutschland

Kremer, Gerhard
gerhard.kremer@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Murr, Sandra
sandra.murr@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Overbeck, Maximilian
maximilian.overbeck@sowi.uni-stuttgart.de
Universität Stuttgart, Deutschland

Pichler, Axel
axel.pichler@ts.uni-stuttgart.de
Universität Stuttgart, Deutschland

Einleitung

Der Workshop zum CRETA Unshared Task (CUTE) verfolgt ein inhaltliches und ein methodisches Ziel. Das inhaltliche Ziel ist die Anregung eines Diskurses über Entitäten, deren Annotation und Kategorisierung entlang von geistes- und sozialwissenschaftlichen Forschungsfragen sowie deren Potential als disziplinübergreifende Textanalyseaufgabe. Methodisch möchten wir ein Workshop-Format erproben, das unseres Erachtens eine produktive Schnittstelle zwischen Geistes-/SozialwissenschaftlerInnen und InformatikerInnen bildet. Das genaue Programm des Workshops wird von den Teilnehmenden durch Beiträge gestaltet (durch Beiträge, siehe Call for Papers [1]) und vor rechtzeitig vor dem Workshop auf der Webseite veröffentlicht [2] .

Entitätenreferenzen

Das Konzept der Entität und ihrer Referenz ist ein bewusst weites, das anschlussfähig sein soll für verschiedene Forschungsfragen aus den Geistes- und Sozialwissenschaften. Wir möchten dabei explizit verschiedene Perspektiven auf Entitäten berücksichtigen.

Entitäten in der Literaturwissenschaft

Figuren in literarischen Texten sind „mit ihrer sinnkonstitutiven und handlungsprogressiven Funktion“ ein zentraler Bestandteil der fiktiven Welt (Platz-Waury 1997). Von besonderem Interesse dabei sind Figurenkonstellationen und Interaktionen, die Entwicklung von Figuren sowie die Funktionalisierung von Figuren als Handlungsträger. Die Erkennung von Figurenreferenzen ist grundlegend, um z.B. Figuren zu charakterisieren, ihre Relationen identifizieren und Netzwerkanalysen durchführen zu können (vgl. Jannidis 2015, Trilcke 2013).

Neben der Figur rückt — spätestens seit dem *spatial turn* — auch der Raum als relevante Entität in den Fokus der Literaturwissenschaft. Der Handlungsraum in literarischen Texten dient der Strukturierung der fiktiven Welt und ist zumeist semantisiert (Lotman 1972). Zudem kann er in Wechselwirkung mit Aspekten der Figur („sujethafte Grenzüberschreitung“, Lotman

1972) oder der Zeit stehen („Chronotopos“, Bachtin 1989).

Entitäten in der Sozialwissenschaft

Politische Parteien, internationale Organisationen oder Institutionen sind seit jeher zentrale Analyseobjekte der empirischen sozialwissenschaftlichen Forschung und werden spätestens seit dem *linguistic turn* (Rorty 1967) in den Sozialwissenschaften auch mittels inhalts- oder diskursanalytischer Methoden auf zunächst kleinen und zunehmend größeren Mengen von Textdokumenten (beispielsweise Parteiprogrammen, offizielle Regierungsdokumenten, Zeitungstexten) untersucht. Neben vielfältigen anderen Analysen stehen dabei oftmals Fragen nach der Sichtbarkeit oder Bewertung bestimmter Entitäten, wie beispielsweise der Europäischen Union als supra-/internationaler Organisation (Kantner 2015) im Vordergrund.

Entitäten in der Philosophie

Im Unterschied zu den Literatur- und Sozialwissenschaften spielen Entitäten als Untersuchungsgegenstand in philosophischen Texten zunächst keine Rolle. Aufgrund ihrer metareflexiven Ausrichtung fragt Philosophie primär nicht nach individuell unterscheidbaren Objekten in der echten oder einer fiktiven Welt, sondern beschäftigt sich mit transzendentalen Fragen nach den Bedingungen und Möglichkeiten derartiger individueller Objekte. Dabei arbeitet sie mit abstrakten Konzepten, die sich ebenfalls als -- nicht-dingliche -- Objekte einer Welt auffassen lassen. Pragmatisch gesehen erfolgt die Referenz auf abstrakte Konzepte in Texten jedenfalls in ähnlicher Weise wie die Referenz auf Figuren, Organisationen und Orten (s.u.).

Fachübergreifende Annotationsschemata

Auch wenn die Interpretation von z.B. der Erwähnung von Organisationen in politischen und des Auftretens von Figuren in literarischen Texten anderen Regeln folgt und mit anderen Forschungsfragen zusammenhängt, gibt es Gemeinsamkeiten auf linguistisch-struktureller Ebene. Im Text realisiert werden Referenzen auf die o.g. Arten von Entitäten entweder als Eigennamen (*Angela Merkel*/ *Ästhetische Theorie*), Pronomen (*sie*/ *sie*) oder als appellative Nominalphrasen (*die Bundeskanzlerin*/ *das Spätwerk Adornos*). Wir haben daher ein einheitliches Vokabular und Annotationsschema entwickelt und auf einem ausgewählten heterogenen Korpus getestet. Dieses soll im Rahmen des Workshops diskursiv erörtert und wenn möglich erweitert werden.

Abstrakt gesprochen verstehen wir unter Entitäten individuell unterscheidbare Objekte in der echten oder einer fiktiven Welt. Wir unterscheiden sechs verschiedene Typen von Entität: Personen, Orte, Ereignisse, Organisationen, kulturelle Artefakte und Konzepte. Die Bezeichnung als „Objekt“ impliziert also *nicht*, dass es sich um physikalische Objekte handelt. Die Einteilung in Typen ist von den oben skizzierten Forschungsfragen und -feldern abgeleitet und ist -- bei anderen Forschungsfragen oder -daten -- offen für Ergänzungen. Die Anwendbarkeit auf zusätzliche Texte und Textgattungen ist für uns (und für diesen Workshop) von besonderem Interesse.

Die Erstellung abstrakter Annotationsrichtlinien und deren systematische, kontrollierte Anwendung (Annotation) auf konkrete Texte verspricht im Wesentlichen zwei Ergebnisse:

Das Erzeugen von parallelen Annotationen auf Basis von Richtlinien zwingt zu einem sehr genauen Lesen des Textes und sorgt für eine intensive Auseinandersetzung mit den Annotationskategorien (und auch für ein Hinterfragen derselben). Recht schnell wird auf diese Weise deutlich, welche Annahmen bei der Anfertigung der Annotationsrichtlinien nicht von den Daten gedeckt waren. Auch Phänomene, die inhaltlich berücksichtigt werden sollten, aber nicht in den Richtlinien enthalten sind, fallen den FachwissenschaftlerInnen schnell ins Auge. Dadurch, dass die eigenen Annotationsentscheidungen ggf. diskutiert und verteidigt werden müssen, sorgen Parallelannotationen für die Aufdeckung von Vagheiten in den Definitionen und damit für eine Klärung der Begriffe (vgl. Gius / Jacke 2016).

Die Entwicklung von maßgeschneiderten Textanalysewerkzeugen für spezifische geistes- und sozialwissenschaftliche Forschungsfragen stößt schnell an

Ressourcengrenzen. Als Problem erweist sich oft, dass die Textanalyseaufgaben zu speziell oder die Datenmengen zu klein sind und damit ein Forschungsbeitrag in der Informatik oder Computerlinguistik nur schwer möglich ist (was typischerweise wiederum Auswirkungen auf den Ressourceneinsatz hat). Eine Antwort auf diese Herausforderung ist die Etablierung fachübergreifender Textanalyseaufgaben, etwa für bestimmte Annotationsebenen. Dies erlaubt die Entwicklung von allgemeineren, wiederverwendbaren Werkzeugen und -- mit geeigneten Testdaten -- deren iterative Verbesserung. Damit wird die Bearbeitung geistes- und sozialwissenschaftlicher Forschungsfragen letztlich nachhaltiger unterstützt als durch die Entwicklung spezieller, aber nach Projektende nicht weiterentwickelter Werkzeuge. Ein Katalysator dafür können shared und unshared tasks sein (vgl. Kuhn / Reiter 2015).

Shared/Unshared Task

In diesem Sinne ist das zweite, methodische Ziel des Workshops zu verstehen: Wir möchten einen Community-Task veranstalten, der eine *shared* und drei *unshared*-Tracks hat. Damit wird ein Workshop-Format auf die Probe gestellt, das eine produktive Schnittstelle zwischen Geistes-/SozialwissenschaftlerInnen und InformatikerInnen zu bilden verspricht (s.a. Belz / Kilgarriff 2006). Im Gegensatz zu *shared tasks*, bei denen die Performanz verschiedener Systeme, Ansätze oder Methoden direkt anhand einer klar definierten und quantitativ evaluierten Aufgabe verglichen wird, sind *unshared tasks* offen für verschiedenartige Beiträge, die auf einer gemeinsamen Datengrundlage oder Fragestellung basieren. Neben dem Call -- der bereits eine Sammlung möglicher Fragestellungen nennt -- veröffentlichen wir daher ein heterogenes Korpus, das als Datengrundlage dient. Im Rahmen von CUTE können Forscherinnen und Forscher an den folgenden Tracks teilnehmen:

Automatische Erkennung von Entitätenreferenzen: Experimente zum automatischen Vorhersagen von Annotationen auf noch nicht annotierten Texten, mit regelbasierten oder statistischen Systemen [3]

Visualisieren von Entitätenreferenzen im Text: Visualisierungsmöglichkeiten zur (interaktiven) Exploration der vorhandenen oder neuen Annotationen

Annotationsanalyse: Qualitative oder quantitative Analyse der vorhandenen Annotationen oder der Annotationsrichtlinien; Annotationsexperimente zur Anwendbarkeit der Richtlinien auf neue Texte

Freestyle: Kreative Ideen, die keinen der obigen Tasks adressieren

Beiträge zu Aufgabe 1 werden quantitativ evaluiert und im Wettbewerb mit den Evaluationsergebnissen der anderen Beiträge verglichen (*shared task*, die technischen Details dazu werden auf der Webseite veröffentlicht). Beiträge für die Aufgaben 2 bis 4 werden vom Programmkomitee qualitativ evaluiert (*unshared task*). Der Austausch während des Workshops (in Form von Kurzvorträgen und Diskussion) wird insoweit eine Bandbreite an Zugängen abbilden, deren verbindendes Element die gemeinsame Datengrundlage sein wird. Da die Teilnehmerinnen und Teilnehmer sich dann im Vorfeld intensiv mit den Daten aus verschiedenen Perspektiven beschäftigen werden, erwarten wir für den Workshop eine erkenntnisreiche Diskussion.

Textgrundlage und Daten

Das von uns im Rahmen des Workshops veröffentlichte Korpus umfasst vier Teilkorpora:

jeweils eine PolitikerInnenrede aus insgesamt vier Parlamentsdebatten des Deutschen Bundestags (S. Leutheuser-Schnarrenberger am 28.10.99, A. Merkel am 16.12.04, A. Ulrich am 15.11.07 und A. Karl am 17.03.11)

Briefe aus Goethes *Die Leiden des jungen Werther* (1787) vom 4. Mai bis einschließlich 16. Juni

der Abschnitt Zur Theorie des Kunstwerks aus Adornos Ästhetische Theorie

die Bücher 3 bis 6 aus Wolframs von Eschenbach *Parzival* (mittelhochdeutsch)

Auch wenn jedes Teilkorpus seine eigenen Besonderheiten hat, wurden alle nach einheitlichen Annotationsrichtlinien annotiert, die wir ebenfalls veröffentlichen und zur Diskussion stellen möchten.

Ausrichter

Der Workshop wird ausgerichtet vom Centre for Reflected Text Analytics (CRETA) an der Universität Stuttgart. CRETA verbindet Literaturwissenschaft, Linguistik, Philosophie und Sozialwissenschaft mit Maschineller Sprachverarbeitung und Visualisierung. Hauptaufgabe von CRETA ist die Entwicklung reflektierter Methoden zur Textanalyse, wobei wir Methoden als Gesamtpaket aus konzeptuellem Rahmen, Annahmen, technischer Implementierung und Interpretationsanleitung verstehen. Methoden sollen also keine "black box" sein, sondern auch für nicht-Technikerinnen und -Techniker so transparent sein, dass ihr reflektierter Einsatz im Hinblick auf geistes- und sozialwissenschaftliche Fragestellungen möglich wird.

Fußnoten

1. http://dhd-blog.org/?p=7333
2. http://www.creta.uni-stuttgart.de/index.php/de/cute/
3. Von dem in der maschinellen Sprachverarbeitung etablierten Task der *named entity recognition* (NER) unterscheidet sich die vorliegende Aufgabe insofern, als dass unsere Annotationen neben Eigennamen auch andere Arten von Referenz enthalten. Werkzeuge (und tasks) zur NER sind darauf getrimmt, ausschließlich Eigennamen zu erkennen.

Bibliographie

Bachtin, Michail Michailowitsch / Kowalski, Edward / Wegner, Michael (1989): *Formen der Zeit im Roman. Untersuchungen zur historischen Poetik*. Frankfurt am Main: Fischer.

Belz, Anja / Kilgarriff, Adam (2006): „Shared-task Evaluations in HLT: Lessons for NLG“, in: *Proceedings of the Fourth International Natural Language Generation Conference*.

Gius, Evelyn / Jacke, Janina (2016): „Kollaboratives Annotieren literarischer Texte“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung*.

Jannidis, Fotis / Krug, Markus / Reger, Isabella / Toepfer, Martin / Weimer, Lukas / Puppe, Frank (2015): „Automatische Erkennung von Figuren in deutschsprachigen Romanen“, in: *DHd 2016: Von Daten zu Erkenntnissen*.

Kantner, Cathleen (2015): *War and Intervention in the Transnational Public Sphere: Problem-solving and European identity-formation*. New York: Routledge.

Kuhn, Jonas / Reiter, Nils (2015): „A Plea for a Method-Driven Agenda in the Digital Humanities“, in: *DH2015: Global Digital Humanities*.

Lotman, Juri (1972): *Die Struktur literarischer Texte*. München: Fink.

Platz-Waury, Elke (1997): „Figur“, in: Weimar, Klaus (ed.): *Reallexikon der deutschen Literaturwissenschaft*. Neubearbeitung des Reallexikon der deutschen Literaturgeschichte. Berlin, New York: de Gruyter 587–589.

Rorty, Richard M. (1967): *The Linguistic Turn*. Chicago: University of Chicago Press.

Trilcke, Peer (2013): „Social Network Analysis als Methode einer textempirischen Literaturwissenschaft“, in: Ajouri, Philip / Mellmann, Katja / Rauen, Christoph (eds.): *Empirie in der Literaturwissenschaft*. Münster: Mentis 201–247.

Daten sammeln, modellieren und durchsuchen mit DARIAH-DE

Gradl, Tobias
tobias.gradl@uni-bamberg.de
Universität Bamberg

Aschauer, Anna
aschauer@ieg-mainz.de
Leibniz-Institut für Europäische Geschichte (IEG)

Dogunke, Swantje
swantje.dogunke@klassik-stiftung.de
Klassik Stiftung Weimar

Klaffki, Lisa
klaffki@hab.de
Herzog August Bibliothek Wolfenbüttel

Schmunk, Stefan
schmunk@sub.uni-goettingen.de
Niersächsische Staats- und Universitätsbibliothek Göttingen

Steyer, Timo
steyer@hab.de
Herzog August Bibliothek Wolfenbüttel

Überblick

Die sammlungsübergreifende Recherche und Nachnutzung geisteswissenschaftlicher Forschungsdaten stehen im Blickpunkt aktueller Forschung in den Digital Humanities. Obwohl das Interesse an einer Zusammenführung digitaler Forschungsdaten bereits kurz nach der Einführung erster digitaler Bibliotheken um die Jahrtausendwende entstand, bleibt die Integration von Forschungsdaten über Sammlungsgrenzen hinweg ein aktuelles Forschungsthema. Bei einer forschungsorientierten Betrachtung von Sammlungen digitaler Daten (also z. B. digitale Texte, Digitalisate, Normdaten, Metadaten) stellt sich die Frage nach den Anforderungen und Erfolgskriterien einer übergreifenden Föderation, Verarbeitung und Visualisierung von Forschungsdaten.

Entgegen der in der Praxis üblichen Orientierung an institutionellen Anforderungen stellen die in DARIAH-DE entwickelten Konzepte und Dienste zur Verzeichnung, Korrelation und Zusammenführung von Forschungsdaten die Bedürfnisse von WissenschaftlerInnen im Kontext ihrer Forschungsfragen in den Mittelpunkt. Dies äußert sich beispielsweise darin, dass DARIAH-DE keine strukturellen Bedingungen an Forschungsdaten stellt. Stattdessen können Daten so publiziert, modelliert und integriert werden, dass eine möglichst gute Passung an den jeweiligen geisteswissenschaftlichen Kontext erreicht wird.

Dieser Workshop wird zunächst in Form kurzer Referate Hintergrundwissen zu den Konzepten und Diensten der DARIAH-DE Föderationsarchitektur [1] vermitteln. Wichtige Bereiche sind dabei nicht nur die Handhabung der Daten selbst sowie Fragen der Lizensierung von Forschungsdaten, sondern auch die Nachnutzbarkeit einmal erhobener oder gesammelter Daten für weitere Forschungsfragen oder zur Nutzung durch andere WissenschaftlerInnen. Ein wesentlicher Anteil des Workshops wird dann insbesondere in der Hands-On-Anwendung der Komponenten durch die TeilnehmerInnen selbst bestehen.

Thematische Schwerpunkte

Die wesentlichen Themenschwerpunkte des Workshops können wie folgt zusammengefasst werden:

- Hintergründe und Best Practices zur *Lizensierung* und *Nachnutzbarkeit* von Forschungsdaten
- Beschreibung und nachhaltige *Referenzierbarkeit* von Sammlungen in der DARIAH-DE Collection Registry
- *Modellierung* von Daten in der DARIAH-DE Schema Registry zur Beschreibung des Erstellungskontexts von Daten sowie deren Transformation in einen forschungsorientierten Verwendungskontext

Anhand der generischen Suche von DARIAH-DE werden die Auswirkungen der Benutzerinteraktion im Rahmen des Workshops sofort erkennbar, d. h. referenzierte Daten werden anhand der entwickelten Datenmodelle verarbeitet und können gemeinsam durchsucht und analysiert werden.

Der gesamte Workshop wird thematisch begleitet von der konkreten, historischen Anforderung (vgl. Szöllösi-Janze, Panter & Paulmann 2015), biographische Daten und Texte aus verschiedenen Datenquellen zu verarbeiten. Die schließlich integrierten biographische Profile (vgl. Gradl & Henrich 2016b) können zur Unterstützung konkreter historischer Forschung herangezogen werden. Das Beispiel ist so gewählt, dass den Teilnehmerinnen und Teilnehmern eine konzeptuelle Übertragung auf ihre eigenen Daten und Forschungsfragen erleichtert wird.

Zielpublikum

Der Workshop richtet sich gleichermaßen an:

- geisteswissenschaftliche WissenschaftlerInnen in den unterschiedlichsten Phasen des akademischen Werdegangs
- VertreterInnen von Institutionen, die Datensammlungen im Rahmen von DARIAH-DE auffindbar und zugreifbar machen möchten,
- sowie auch VertreterInnen der Informatik, die ein Interesse an der Implementierung von DARIAH-DE Komponenten bzw. den

Datenaustausch auf Basis maschinell zugreifbarer Schnittstellen haben.

Wer bereits über digitale Daten verfügt, ist herzlich eingeladen, diese für die Hands-On-Sessions mitzubringen, um an diesen konkreten Beispielen die DARIAH-DE-Tools zu erarbeiten. Für TeilnehmerInnen, die keine geeigneten Daten mitbringen können, werden Beispiele zur Verfügung gestellt. Bitte bringen Sie in jedem Fall Ihren eigenen Laptop mit!

Der Workshop ist Teil zweier konzeptionell eigenständiger Einreichungen zu den Angeboten der digitalen Forschungsinfrastrukturen TextGrid und DARIAH-DE. Der erste Workshop hat den Titel "Annotieren und Publizieren mit DARIAH-DE und TextGrid“. Der Besuch beider Workshops ermöglicht eine grundlegende und umfassende Einführung in und Anwendung von Architektur, Tools, Diensten und Workflows zum Annotieren, Sammeln, Modellieren, Recherchieren und Publizieren geisteswissenschaftlicher Forschungsdaten.

Inhalte und Ablauf des Workshops

I - Impulsreferate “Sammeln”

- Lizensierung, Referenzierung und Nachnutzbarkeit von Forschungsdaten (*Lisa Klaffki*)
- Transnationale Biographien als Beispiel einer historischen Motivation für die forschungsorientierte Föderation von DARIAH-DE (*Anna Aschauer*)

II - Impulsreferat “Modellieren”

- Forschungsorienterte Modellierung und Korrelation von Daten in der Föderationsarchitektur von DARIAH-DE (*Stefan Schmunk, Tobias Gradl*)

III - Impulsreferat “Durchsuchen”

- Integriertes Suche über heterogene Datenbestände – Anforderungen und Lösungsansätze im Bereich des kulturellen Erbes (*Timo Steyer, Swantje Dogunke*)

IV - Hands-on Session “Sammeln, Modellieren & Durchsuchen”

- Anwendung der Föderationsarchitektur und generischen Suche von DARIAH-DE (*Tobias Gradl*)
  - Modellierung von Daten und Vorbereitung einer Nachnutzung
  - Assoziation heterogener wissenschaftlicher Sammlungen
  - Verfeinerung der benutzerdefinierten Suchmöglichkeiten in der generischen Suche (Suchbild, Ranking etc.)
  - Anpassung der generischen Suche und Bereitstellung benutzerdefinierter Suchen

Komponenten des Workshops

Abbildung 1 zeigt die Zusammenhänge zwischen den für die DARIAH-DE Infrastruktur zugänglichen Kollektionen, den Registries und der generischen Suche. In der Übersicht dargestellte Komponenten und Verbindungen werden im Rahmen des Workshops live durch die TeilnehmerInnen beeinflusst, weshalb wir in diesem Abschnitt eine vorbereitende Einführung anbieten möchten. Für weitere Informationen erlauben wir uns einen Verweis auf die weiterführenden Publikationen am Ende des Dokuments.

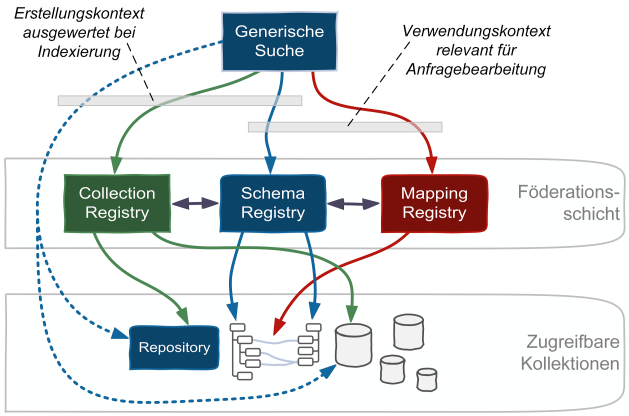


Abbildung 1: DARIAH-DE Föderationsarchitektur

“Sammeln”: Collection Registry

Die Collection Registry (vgl. Abbildung 2) ist ein zentrales Verzeichnis zur Registrierung und Beschreibung von Sammlungen von Ressourcen. Sammlungen können selbst direkt Ressourcen oder weitere untergeordnete Teilsammlungen beinhalten und können sowohl physische als auch digitale Objekte oder nur Daten aggregieren. Die Sammlungsbeschreibungen

decken neben Verschlagwortung, zeitlichen und geografischen Dimensionen auch Sammlungsformate und Informationen zur Datenpflege ab.

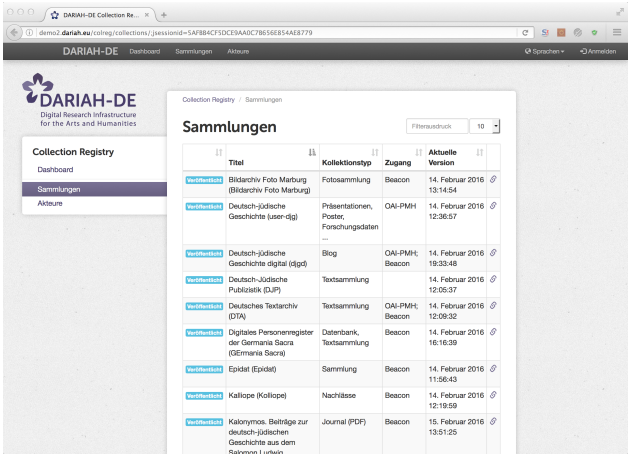


Abbildung 2: Bildschirmausschnitt „Sammlungen“ der DARIAH-DE Collection Registry

“Modellieren”: Schema Registry / Mapping Registry

In der Schema und Mapping Registry werden Datenmodelle und Korrelationen zwischen diesen beschrieben. Die grundlegende Zielsetzung besteht in der Definition und nachnutzbaren Modellierung der Erstellungs- und Verwendungskontexte von Daten:

- *Erstellungskontext*: Ausgehend beispielsweise von einem XML-Schema wird ein Datenmodell angelegt, verfeinert und um Hintergrundwissen z. B. zur Sammlung, Institution erweitert (vgl. Abbildung 3). Hierdurch wird insbesondere eine Nachnutzung von Daten außerhalb des originären Sammlungskontexts ermöglicht.
- *Verwendungskontext*: Durch die Definition eines fallspezifischen Integrationsmodells können Datenmodelle miteinander assoziiert werden. Durch eine Formulierung von Transformationsregeln werden Daten so umgewandelt und integriert, wie sie für eine weiterführende Untersuchung benötigt werden (vgl. Abbildung 4).

Abbildung 3: Bildschirmausschnitt des Schema Editors

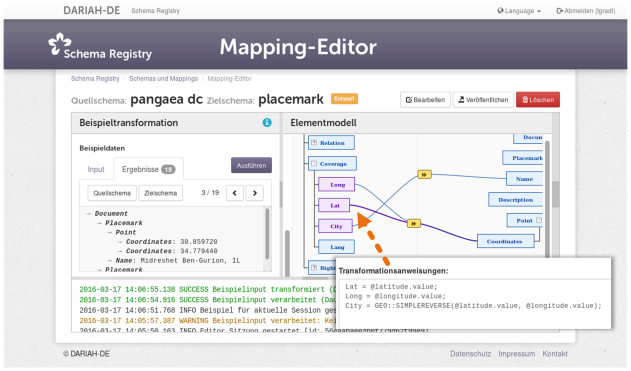


Abbildung 4: Bildschirmausschnitt des Mapping Editors

“Durchsuchen”: Generische Suche

Mit der generischen Suche wird im Rahmen von DARIAH-DE ein konkreter Anwendungsfall der Datenföderation umgesetzt. Hierbei werden Daten aus den in der Collection Registry verzeichneten Kollektionen nach den in der Schema Registry definierten Datenmodellen verarbeitet und indexiert. Die Heterogenität der Ressourcen wird zum Zeitpunkt konkreter Suchanfragen, basierend auf der zu durchsuchenden Menge von Kollektionen, mit Hilfe der Mapping Registry aufgelöst.

Über die Möglichkeit der einfachen Suche über die Daten verzeichneter Kollektionen hinaus, können auf Basis der Funktionalität der generischen Suche weiterführende, fachspezifische Suchmaschinen implementiert werden (s. Abbildung 5).

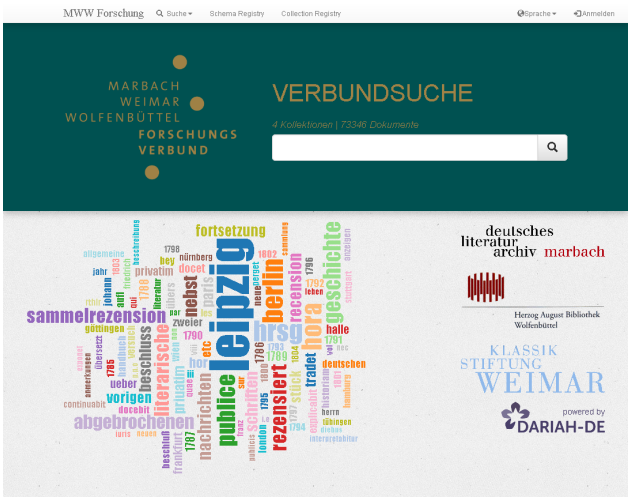


Abbildung 5: Fachwissenschaftliche Spezialsuche im Rahmen der generischen Suche

"Durchsuchen": Historischer Use-Case Biographien

Der Use-case *Biographien* verbildlicht wie man eine historische Fragestellung anhand digitaler Werkzeuge bearbeiten kann.

Prosopographische historische Forschung orientiert sich immer noch stark an nationaler Geschichtsschreibung: religiöse, berufliche, gesellschaftliche Gruppen werden oft innerhalb der nationalen Grenzen, die selbst ein Konstrukt der Moderne sind, untersucht. Das Zusammenführen der Daten aus unterschiedlichen biographischen Datenbanken kann helfen dieses Problem zu lösen und biographische Recherchen über die nationalen Grenzen hinweg zu gestalten.

Zu diesem Zweck implementiert DARIAH-DE derzeit das CosmoTool (vgl. Gradl & Henrich 2016b), welches auf die Unterstützung historischer Forschung an biographischen Daten abzielt. Das Werkzeug kann dabei als logische Konsequenz einer Spezialisierung der generischen Suche interpretiert werden:

- die Sammlung von Datenquellen erfolgt in der DARIAH-DE Collection Registry,
- die Modellierung der Daten, sowie deren Assoziation mit einem zentralen, biographischen Schema erfolgt in der DARIAH-DE Schema / Mapping Registry
- die Verarbeitung und Indexierung der Daten basiert auf funktionalen Komponenten der generischen Suche
- Die Analyse und Visualisierung wurde und wird dagegen spezifisch für den Anwendungsfall entwickelt und bildet den tatsächlichen Kern des CosmoTools

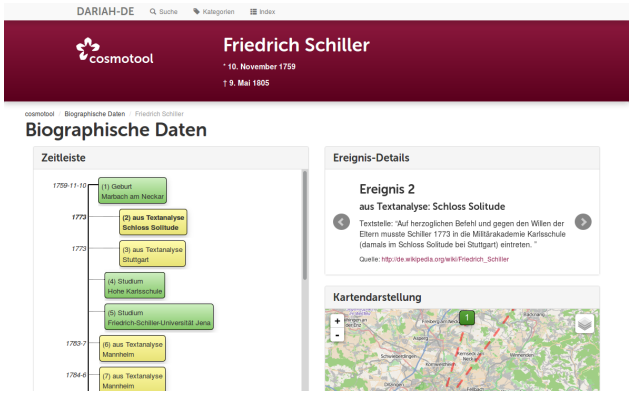


Abbildung 6: Bildschirmausschnitt des CosmoTools

Zusammenfassung

Insgesamt werden den TeilnehmerInnen im Rahmen dieses Workshops verschiedene Kenntnisse im Kontext der Sammlung, Modellierung und Suche geisteswissenschaftlicher Forschungsdaten vermittelt. Durch die Anwendung der entsprechenden Komponenten von DARIAH-DE werden die in vorausgegangenen Referaten vorgestellten Ideen vertieft.

Die Begleitung des Workshops durch Forschungsfragen und Daten im Kontext biographischer Daten soll den TeilnehmerInnen die praktische Anwendung der Komponenten deutlich machen. Idealerweise wird dadurch die Übertragbarkeit auf andere Daten und Fragen vermittelt, wodurch eine nachhaltige Zugänglichkeit wissenschaftlicher Forschungsdaten erreicht werden kann.

Kontaktdaten aller Beitragenden

Anna Aschauer, Leibniz-Institut für Europäische Geschichte (IEG), Querschnittsbereich, Alte Universitätstraße 19, 55116 Mainz
aschauer@ieg-mainz.de
Forschungsinteressen: Pietismusforschung, Geschichte Russlands, Migration der religiösen Minderheiten in der Frühen Neuzeit, Digital Humanities.

Swantje Dogunke, Forschungsverbund Marbach Weimar Wolfenbüttel / Klassik Stiftung

Weimar, Direktion Verwaltung, Abteilung Informationstechnik, Burgplatz 4, 99423 Weimar
swantje.dogunke@klassik-stiftung.de
Forschungsinteressen: Dokumentation im Museum, Museumsmanagement, digital curation, digitale Langzeitarchivierung, Digial Humanities

Tobias Gradl, Otto-Friedrich-Universität Bamberg, Lehrstuhl für Medieninformatik, An der Weberei 5, 96052 Bamberg
tobias.gradl@uni-bamberg.de
Forschungsinteressen: Forschungsdaten und Forschungsdatenmanagement, Digital Humanities, Datenintegration, Information Retrieval

Lisa Klaffki, Herzog August Bibliothek Wolfenbüttel, Abteilung 1, Lessingplatz 1, 38304 Wolfenbüttel
klaffki@hab.de
Forschungsinteressen: Archäologie der germanischen Provinzen, Bestattungssitten der römischen Kaiserzeit, Digital Humanities

Stefan Schmunk, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Abt. Forschung und Entwicklung, Papendiek 14, 37073 Göttingen,
schmunk@sub.uni-goettingen.de
Forschungsinteressen: Forschungsdaten und Forschungsdatenmanagement, Digitale Geschichtswissenschaft, Virtuelle Forschungsumgebungen, Digitale Forschungsinfrastrukturen

Timo Steyer, Forschungsverbund Marbach Weimar Wolfenbüttel / Herzog August Bibliothek Wolfenbüttel, Abteilung 1, Lessingplatz 1, 38304 Wolfenbüttel
steyer@hab.de
Forschungsinteressen: Digitale Editionen, Datenmodellierung und Metadaten, Digital Humanities

Zahl der möglichen Teilnehmerinnen und Teilnehmer

Die Zahl der möglichen Teilnehmer ist aus unserer Sicht nicht eingeschränkt. Einer sehr großen Zahl müsste ggf. durch mehrere Helfer in der Hands-On-Session entgegnet werden

Angaben zu einer etwa benötigten technischen Ausstattung

Es wird keine zusätzliche Ausstattung neben der üblichen Präsentationstechnik benötigt. Von den TeilnehmerInnen wird das Mitbringen eines eigenen Laptops für die aktive Teilnahme an der Hands-On-Session erwartet.

Fußnoten

1. Repository, Collection Registry, Schema / Mapping Registry und Generische Suche von DARIAH-DE (vgl. Gradl & Henrich 2016a, Schmunk & Funk 2016)

Bibliographie

Gradl, Tobias / Henrich, Andreas (2016a): „Die DARIAH-DE Föderationsarchitektur - Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen“, in: *Bibliothek - Forschung und Praxis 2016* 40 (2): 222–228 10.1515/bfp-2016-0027.

Gradl, Tobias / Henrich, Andreas (2016b): „Nutzung und Kombination von Daten aus strukturierten und unstrukturierten Quellen zur Identifikation transnationaler Lebensläufe“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 129–132.

Gradl, Tobias / Lordick, Harald / Henrich, Andreas (2016): „Judaica recherchieren: Unterstützung bei der Realisierung forschungsspezifischer Suchlösungen durch die generische Suche“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 132–136.

Schmunk, Stefan / Funk Stefan (2016): „Das DARIAH-DE- und das TextGrid-Repositorium: Geistes- und kulturwissenschaftliche Forschungsdaten persistent und referenzierbar langzeitspeichern“, in: *Bibliothek - Forschung und Praxis 2016* 40 (2): 213–221 10.1515/bfp-2016-0020.

Szöllösi-Janze, Margit / Panter, Sarah / Paulmann, Johannes (2015): „Mobility and Biography. Methodological Challenges and Perspectives“, in: *Jahrbuch für Europäische Geschichte / European History Yearbook* 16: 1–14 10.1515/9783110415162-001.

Dokumente segmentieren und Handschriften erkennen: Arbeiten mit der Plattform Transkribus

Hodel, Tobias
tobias.hodel@hist.uzh.ch
Staatsarchiv des Kantons Zürich

Lang, Eva-Maria
Eva.Lang@bistum-passau.de
Archiv des Bistums Passau

Fiel, Stefan
fiel@caa.tuwien.ac.at
Technische Universität Wien, Faculty of Informatics, Institute of Computer Aided Automation, Computer Vision Lab

Die Aufbereitung und Erkennung von handschriftlichen Dokumenten ist sowohl für Menschen als auch für Computeralgorithmen eine technische Herausforderung. Die Bearbeitung von handschriftlichem Material wird bislang von spezialisierten Experten durchgeführt, um technisch und qualitativ hochstehende Resultate aus historischen Dokumenten zu erhalten. Zur Erstellung hochwertiger Editionen ist dafür hilfswissenschaftliches Wissen (Paläographie, Editorik), historisches Hintergrundwissen und technisches Know-how gefragt.

Im Rahmen des Projekts READ (Recognition and Enrichment of Archival Data) werden unterschiedliche Aufgaben der Automatisierung (weiter-)entwickelt, um qualitativ gute Ergebnisse mit optimalem Ressourceneinsatz zu erhalten. Ein speziell dafür entwickeltes Tool ist die Software Transkribus, die die Arbeit von Experten und maschineller Erkennleistung verkoppelt. Die Software ist frei verfügbar unter www.transkribus.eu. Im Workshop wird Transkribus vorgestellt und kann durch die Teilnehmenden mit eigenen oder zur Verfügung gestellten Dokumenten getestet werden.

Transkribus unterstützt alle Prozesse vom Import der Bilder über die Identifikation der Textblöcke und Zeilen, die zu einer detaillierten Verlinkung zwischen Text und Bild führt sowie die Transkription und Annotation der Handschrift bis zum Export der gewonnen Daten in standardisierten Formaten.

Workflow in Transkribus

Um Texte zu transkribieren oder zu edieren, müssen digitale Bilder hochgeladen und danach mit Layouterkennungswerkzeugen bearbeitet werden. Die Analyse des Layouts kann automatisiert geschehen, wobei die manuelle Kontrolle und falls nötig die Nachbearbeitung im Moment noch sinnvoll ist.

Texte aus in Transkribus aufbereiteten Dokumenten können entweder mit bereits bestehenden HTR-Modellen (Handwritten Text Recognition) erkannt oder händisch erstellt werden und danach zum Training neuer Modelle genutzt werden. Insbesondere für die Bearbeitung grosser Dokumentenkorpora, die in ähnlichen Handschriften verfasst wurden, lassen sich bereits heute Effizienzgewinne und Vereinfachungen erzielen.

Aufbauend auf den Transkriptionen ist es möglich eine Vielzahl von Auszeichnungen und Annotationen innerhalb des Textes, aber auch darüber hinaus für Einzeldokumente und ganze Dokumentenbestände anzulegen. Neben der Anreicherung der Dokumente mit der Identifikation von Personen, Orten und Sachwörtern ist somit auch die Möglichkeit der Herstellung von Bestandsbeschreibungen und der Hinterlegung von Transkriptions- und Editionsvorschriften gegeben.

Ausgabeformate

Für den Export stehen unterschiedliche Formate und Ausgabeformen zur Verfügung. So ist es möglich XML-Dateien zu exportieren, die den Vorgaben der TEI entsprechen. Ausgehend davon können komplexe digitale Editionen erstellt werden, die jedoch im Unterschied zu herkömmlichen Editionen eine enge Verzahnung mit den verwendeten Bilddateien aufweisen. Dadurch werden Editionen ermöglicht, die den transkribierten Text in der Zusammenschau mit der faksimilierten Vorlage sichtbar machen (analog zu state-of-the-art Editionen, wie beispielsweise die Edition der Briefe Alfred Eschers: https://www.briefedition.alfred-escher.ch/). Daneben sind auch Ausgaben als Druckdaten (PDF) oder zur Weiterbearbeitung für Textverarbeitungsprogramme (DOCX) implementiert. Schliesslich ist auch ein Export im PAGE-Format (zur Anzeige in Viewern für OCR gelesene Dokumente, Pletschacher, 2010) sowie als METS (Metadata Encoding and Transmission) möglich.

Die Speicherung der Dokumente erfolgt in der Cloud (gehostet auf Servern der Universität Innsbruck). Die importierten Daten bleiben auch während der Bearbeitung unverändert

im Dateisystem liegen und werden ergänzt durch METS und PAGE XML, letztere in eigenem Unterordner. Alle bearbeiteten Dokumente und Daten bleiben somit in den unterschiedlichen Bearbeitungsstadien nicht nur lokal verfügbar, sondern können für Projektmitarbeitende geteilt werden. Dank elaboriertem *user-management* ist die Zuteilung von Rollen möglich. Die Erkennprozesse werden serverseitig durchgeführt, sodass die Ressourcen auf den lokalen Rechnern nicht strapaziert werden. Transkribus ist mit JAVA und SWT programmiert und kann daher plattformunabhängig (Windows, Mac, Linux) genutzt werden.

Zielpublikum

Die Plattform ist für unterschiedliche Gruppen konzipiert. Einerseits für GeisteswissenschaftlerInnen, die selbst Transkriptionen und Editionen historischer Dokumente erstellen möchten. Andererseits richtet sich die Plattform an Archive, Bibliotheken und andere Erinnerungsinstitutionen, die handschriftliche Dokumente in ihren Sammlungen aufbewahren und ein Interesse an der Aufbereitung des Materials haben. Angesprochen werden sollen auch Studierende der Geistes-, Archiv- und Bibliothekswissenschaften mit einem Interesse an der Transkription historischer Handschriften.

Das Ziel, eine robuste und technisch hochstehende Automatisierung von Layout und Handschrift, lässt sich nur durch die enge Zusammenarbeit zwischen Geisteswissenschaftlern und Computerspezialisten erreichen, die bezüglich Datenqualität und Herstellung von Transkriptionen von unterschiedlichen Voraussetzungen und Ansprüchen ausgehen. Die Algorithmen werden daher nicht nur bis zu einem Status als *proof-of-concept* erarbeitet, sondern bis zur Praxistauglichkeit verfeinert und in grösseren Forschungs- und Aufbewahrungsumgebungen getestet und verbessert. Die Computerwissenschaftler sind entsprechend ebenfalls ein wichtiges Zielpublikum, wobei bei ihnen weniger die Nutzung der Plattform als das Beisteuern von Software(teilen) anvisiert wird. Die eingespeisten Dokumente und Daten bleiben privat und vor dem Zugriff Dritter geschützt. Von Projektseite können vorgenommene Arbeitsschritte zwecks besserem Verständnis der ausgeführten Arbeiten und letztlich der Verbesserung der Produkte ausgewertet werden.

Layout- und Texterkennung

Die zwei zentralen Automatisierungsprozesse basieren auf Algorithmen, die in laufenden Forschungsprojekten entwickelt und verbessert werden. Die *document image analysis* (DIA) versucht Textblöcke zu identifizieren und von Dreck, Scanfehlern und anderen Störsignalen zu unterscheiden, wobei zwischen handschriftlichen und gedruckten Textblöcken differenziert wird (Zagoris 2012; Stamatopoulos 2015).

In Transkribus werden auf der Layouterkennung aufbauend zwei *handwritten text recognition*-Engines (HTR) angeboten, die auf unterschiedlichen technischen Grundlagen basieren: Erstens kann eine nach dem Hidden Markov Model (HMM) operierende HTR der Technischen Universität Valencia angewählt werden (Toselli 2015, Puigcerver 2015). Zweitens kann ein Model basierend auf rekurrierenden neuronalen Netzwerken der Universität Rostock genutzt werden (Leifert 2016).

Transkribus und das gesamte Forschungsnetzwerk will die verfügbaren technischen Möglichkeiten den Endnutzern nach möglichst gängigen Workflows aufbereiten, so dass dem schnellen Praxiseinsatz keine Hindernisse im Weg stehen. Im Gegenzug wird die Nutzung im grossen Umfang erhofft, die den Subprojekten wichtige Trainingsdaten und Aufschlüsse bezüglich der Nutzung und den Problemen mit den Algorithmen sowie dem Graphical User Interface geben. Tests zum Einsatz der Technik in Archiven und Bibliotheken und unter unterschiedlichen Bedingungen werden momentan getestet und evaluiert.

Als Businessmodel ist eine Überführung des Forschungsprojekts in eine Kooperative geplant, die den Stakeholdern möglichst niederschwellige und kostengünstige Angebote unterbreiten soll (Mühlberger, Preprint). Somit vereint das Projekt READ die unterschiedlichsten Ansprüche an Automatisierungs- und Erkennungsroutinen und orientiert sich dabei an gängigen Arbeitsformen im Kontext mit handschriftlichen Dokumenten (siehe auch die Projekthomepage: http://read.transkribus.eu).

Aus- und Seitenblicke im Workshop

Zwei unterschiedliche Forschungsaspekte aus READ werden im Rahmen des Workshops als Inputs demonstriert:

Einerseits der Umgang mit einer speziellen Dokumentenform, Kirchenbüchern, in denen stark strukturierte Daten aus Pfarreien gesammelt wurden (Wurster, 2014 / 2015). Aufgrund der Strukturerkennung und der HTR wird es möglich, spezialisierte Suchroutinen zu produzieren.

Andererseits können aufgrund der erhobenen Daten und durch c *omputer vision* Profile

der Schreibenden erstellt werden, die die Identifikation der Personen als Schreibende weiterer Dokumente naheliegend macht (Fiel, 2012). Beide Anwendungen versprechen für die Geisteswissenschaften neue Zugänge zu grossen Datenschätzen, die in den handschriftlichen Beständen gehoben werden können.

Programm/Ablauf des Workshops

Einführung in Transkribus (Tobias Hodel, Zürich): 30‘
Aufbau und Funktionieren des Programms, Demonstration des Gebrauchs anhand von Beispielen. Aufzeigen der Möglichkeiten zum Einsatz der Automatisierungen.

Strukturierte Daten in Kirchenbüchern (Eva-Maria Lang, Passau): 30‘
Demonstration vom Umgang mit Kirchenbüchern, einer spezifischen und stark standardisierten Dokumentform, die mit Transkribus aufbereitet werden. Eine Suche in den Dokumenten wird über eigene Routinen und Abfragemöglichkeiten gewährleistet.

Selbstständiges Arbeiten der Teilnehmenden mit Transkribus: 90‘
Die Möglichkeiten und Grenzen von Transkribus sollen von den Teilnehmenden (wenn möglich mit eigenen Dokumenten) selbst ausgetestet werden.

Schreiberidentifizierung (Stefan Fiel, Wien): 30‘
Ein über Transkribus hinausgehender Teil des Projekts beschäftigt sich mit *computer vision*. Ziel ist die Identifizierung unterschiedlicher Personen als Schreibende. Stefan Fiel berichtet über den Stand der Forschung und wie Teilnehmende die Hände wichtiger Schreibender zur Verfügung stellen können.

Diskussion über Vor- und Nachteile der Software: 45‘
Inklusive Evaluation des Tools und der Veranstaltung. Feedbacks werden eingeholt, zur Verbesserung der Software (usability, Umfang und Leistung der Automatisierungen etc.).

Das Projekt READ und somit die Weiterentwicklung von Transkribus werden finanziert durch einen Grant der Europäischen Union im Rahmen des Horizon 2020 Forschungs- und Innovationsprogramms (grant agreement No 674943).

Kontaktdaten aller Beitragenden (inkl. Forschungsinteressen)

Tobias Hodel, Staatsarchiv des Kantons Zürich, Winterthurerstrasse 170, CH-8057 Zürich, Schweiz; tobias.hodel@ji.zh.ch (Digital Humanities; Handwritten Textrecognition; eArchiving; Information Retrieval).

Eva-Maria Lang, Archiv des Bistums Passau, Luragogasse 4, DE-94032 Passau, Eva.Lang@bistum-passau.de (Automatic Text Recognition, Digital Archives, Image Recognition and Information Retrieval, Software Architecture).

Stefan Fiel, Technische Universität Wien, Faculty of Informatics
Institute of Computer Aided Automation, Computer Vision Lab, Favoritenstr. 9/183-2, A-1040 Vienna, Austria; fiel@caa.tuwien.ac.at (Bilderverarbeitung und Dokumentenanalyse).

Zahl der möglichen Teilnehmerinnen und Teilnehmer

30-40 Personen (auch abhängig von der Raumgrösse)

Benötigte technische Ausstattung:
Allgemein: Beamer, evtl. Whiteboard.
Teilnehmende: Eigener Rechner (wenn möglich Installation von Transkribus; Hilfe zur Installation von Transkribus wird 30 Minuten vor der Veranstaltung angeboten)

Anmeldungen und Rückfragen bitte an tobias.hodel@ji.zh.ch

Bibliographie

Fiel, Stefan / **Sablatnig, Robert** (2012): „Writer Retrieval and Writer Identification using Local Features“, in: *10th IAPR International Workshop on Document Analysis Systems* http://www.ict.griffith.edu.au/das2012/attachments/FullPaperProceedings/4661a145.pdf .

Leifert, Gundram / **Strauß, Tobias** / **Grüning, Tobias** / **Labahn, Roger** (2016): *Cells in Multidimensional Recurrent Neural Networks* https://arXiv.org/abs/1412.2620v02 .

Mühlberger, Günter / **Colutto, Sebastian** / **Kahle, Philip** (Preprint): *Handwritten Text Recognition (HTR) of Historical Documents as a Shared Task for Archivists, Computer Scientists and Humanities Scholars: The Model of a Transcription & Recognition Platform (TRP).*

Pletschacher, Stefan / **Antonacopoulos, Apostolos** (2010): „The PAGE (page analysis and ground-truth elements) format framework“, in: *Proc. ICPR* 257–260.

Puigcerver, Joan / **Toselli, Alejandro Héctor** / **Vidal, Enrique** (2015): „Probabilistic interpretation and improvements to the hmm-filler for handwritten keyword spotting“, in: *13th international conference on document analysis and recognition (ICDAR).*

Stamatopoulos, Nikolaos / **Gatos, Basilis** (2015): „Goal-oriented performance evaluation methodology for page segmentation techniques“, in: *13th international conference on document analysis and recognition (ICDAR)* 281–285.

Toselli, Alejandro Héctor / Vidal, Enrique (2015): „Handwritten text recognition results on the Bentham collection with improved classical n-gram-HMM methods“, in: *International workshop on historical document imaging and processing (HIP)*.

Wurster, Herbert W. (2015): „Schritt für Schritt ins Internet – Europas Matriken online“, in: *insights: Archives and people in the digital age* 2: 16–17.

Wurster, Herbert W. (2014): „Matrikeln - Ein kulturhistorischer Blick auf die Kirchenbücher“, in: *Zeitschrift für bayerische Kirchengeschichte* 83: 87–93.

Zagoris, Konstantinos / Pratikakis, Ioannis / Antonacopoulos, Apostolos / Gatos, Basilis / Papamarkos, Nikos (2012): „Handwritten and Machine Printed Text Separation in Document Images Using the Bag of Visual Words Paradigm“, in: *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference* 103–108 10.1109/ICFHR.2012.207.

Einführung in das PANDORA Linked Open Data Framework.

Johnson, Christopher

christopher.johnson@uni-goettingen.de
Akademie der Wissenschaften zu Göttingen, Deutschland

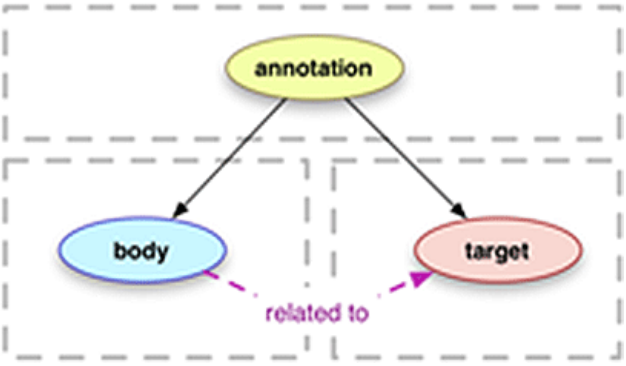
Wettlaufer, Jörg

jwettla@gwdg.de
Akademie der Wissenschaften zu Göttingen, Deutschland

Beschreibung des Workshops [Zeitrahmen 4h]

Der Workshop stellt eine Softwarearchitektur vor, die zurzeit im Rahmen des Projekts „Johann Friedrich Blumenbach – online“ der Göttinger Akademie der Wissenschaften im Zusammenhang mit der geplanten digitalen Edition der gedruckten Werke und naturhistorischen Sammlungen J.F. Blumenbachs (1752-1840) entwickelt wird. Bei der Konzeption stehen Interoperabilität, Erweiterbarkeit und Nachnutzung als zentrale Entwicklungsziele im Vordergrund. Ausgangspunkt des **PANDORA** [**P**resentation (of) **AN**notations (in a) **D**igital **O**bject **R**epository **A**rchitecture] Linked Open Data (LOD) Frameworks sind digitale Abbildungen von Texten und Objekten, die in einem Fedora Commons Repository(1) gespeichert und über das International Image Interoperability Framework (IIIF) visualisiert werden. Das Framework ist insbesondere für den Einsatz im Museumskontext und im Bereich der digitalen Präsentation von Kulturgutüberlieferung geeignet. Dabei können sowohl text- also auch objektbasierte Fragestellungen untersucht bzw. Kulturgüter präsentiert und digital verfügbar gemacht werden. Ein besonderer Vorteil ist dabei die Bereitstellung der Daten als LOD und die Möglichkeit der Einbindung der Ressourcen in andere Kontexte. In dem Workshop sollen die Einsatz- und Nachnutzungsmöglichkeiten sowie die Nachhaltigkeit dieser Architektur vorgestellt, diskutiert und anhand von Beispielanwendungen zusammen mit den Teilnehmerinnen und Teilnehmern erprobt werden.

PANDORA ist zunächst einmal eine Sammlung von Open Source Anwendungen, die über ein gemeinsames „Manifest“ Dokument die Präsentation der Daten für den Anwender organisieren. Das „Manifest“ besteht aus einem JSON-LD(2) Dokument und wird aus einem digitalen Objektrepositorium über die dynamische Verwendung von SPARQL-Abfragen(3) erzeugt. Es orientiert sich dabei an der Semantik und dem Konzept der „IIIF Presentation API“(4). Diese Schnittstelle definiert, wie die Struktur und das Layout eines komplexen und bild-basierten Objekts in einem Standardformat dargestellt werden kann und zielt darauf ab, die Interoperabilität und Erweiterbarkeit von Präsentationen basierend auf dem Open Annotation Datenmodell(5) zu erleichtern. In diesem Modell ist oa:Annotation jede Ressource, die aus zwei Komponenten besteht, einen „body“ und einen „target“:



[Abb. 1: Annotation Datenmodell]

In der IIIF Presentation API ist das Ziel ein "canvas" (eine Leinwand), der eine Abstraktion des Client-Arbeitsplatz oder Sichtbereichs darstellt. Die Annotation (body) kann mit jedem verknüpften oder eingebetteten Objekt wie einem Bild, einer Beschreibung oder einem semantischen Tag verlinkt sein. Die assoziative Beziehungen zwischen verschiedenen Annotation-„bodies“ auf einem „canvas“ sind mit der Linked-Data Semantik im Manifest instanziiert. Die Segmentierung ermöglicht die Auswahl eines Bereichs eines Bildes oder eines „canvas“ unter Verwendung rechteckiger Begrenzungsrahmen oder mit der „IIIF Image API“(6), einem „stream“ von Bildausschnitten. Hotspot Verknüpfungen ermöglichen es die Auswahl auf ein Anmerkungsobjekt zu lenken, um eine Zustandsänderung in einem anderen Annotationsobjekt auszulösen.

Die Annotationen existieren im Fedora-Repository als LDP Container(7), der in einer Hierarchie von Ressourcen eine HTTP-adressierbare Ressource ist. Wenn der LDP Container in einem Triple-Store überführt wird, existiert er dort als RDF Ressource und als sog. „Named Graph“(8). Der IIIF Manifest Service unterstützt die Serialisierung bzw. „Kanonikalisierung“(9) des JSON-LD Dokuments in Form einer geordneten Liste, die im Ressource Description Framework als „collection" bezeichnet wird. Die Darstellung einer Manifest-Sequenz eines „canvas“ als RDF Sammlung erfordert die Verwendung von leeren Knoten, sog. „blank nodes“, die wie folgt miteinander verwoben sind:

<LDP_Manifest_Sequence_Container> sc:hasCanvases _:c11 .

_:c11 rdf:first <http://localhost:8080/fcrepo/rest/edition/base/canvas/c000> .

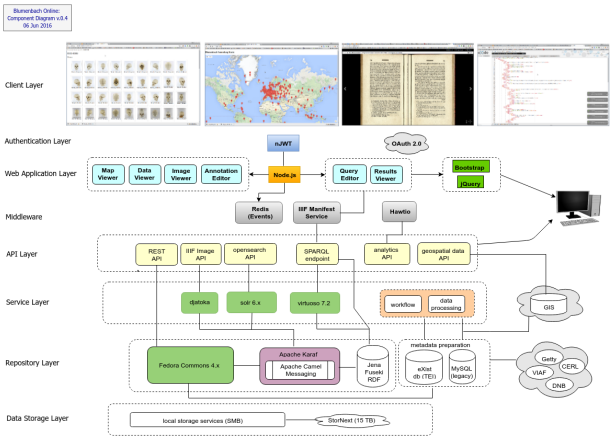
_:c11 rdf:rest _:c001.

_:c001 rdf:first <http://localhost:8080/fcrepo/rest/edition/base/canvas/c001> .

_:c001 rdf:rest ...

Im Fedora-Repository wird ein „blank node“ mit einer bekannten Skolem IRI [nach RFC5785(10)] repräsentiert.

Durch die Verwendung des PANDORA IIIF Manifest Services(11) wird die Konstruktion von Präsentationen aus SPARQL Abfragen erlaubt, die eine sehr differenzierte Darstellung der Annotationen über JSON-LD ermöglichen. Der Entwurf einer LDP Container-Hierarchie und von Sammlungs-Definitionen im Einklang mit der Semantik der IIIF Presentation API "Annotation-Liste"(12) und "Layer"(13) für die Darstellung von Textsequenzen (Zeilen, Wortgruppen, Absätze, Seiten, Kapitel, etc.) ist ein integraler Bestandteil von PANDORA. Das folgende Schaubild verdeutlicht die Architektur des Frameworks und die Verknüpfung und das Zusammenspiel der einzelnen Komponenten:



[Abb. 2: PANDORA Architektur]

Mit einer klaren Trennung der Domain- und Client-Rollen bietet das PANDORA Framework Flexibilität und Erweiterbarkeit für alle möglichen Web-Client Präsentationsmethoden. Darüber hinaus unterstützt PANDORA Node.js Instanzen, die durch socket.io und Redis Pub/Sub(14) Ereignisse verbunden sind und dadurch Redundanz und Durchsatz für dezentrale asynchrone Operationen bieten. Das Framework besteht aus aktueller Open Source Software nach Industriestandards für Linked Data. Dazu gehören das Fedora-Repository, Apache Jena, Apache Camel, Apache Karaf, Open Virtuoso und Solr. Es ist gekennzeichnet durch Interoperabilität, Flexibilität und Erweiterbarkeit und erlaubt, durch die Verwendung von Standard-Software, ebenfalls eine Nachnutzung der Forschungsdaten über Linked Open Data Schnittstellen. Diese Daten können über den SPARQL-Endpoint entweder lokal integriert oder extern zur Nachnutzung angeboten werden. Weitere Informationen finden sich im GitHub Repository.(15) Eine ausführliche Dokumentation sowie eine Webseite mit Links zum Download der Komponenten befinden sich in Vorbereitung.

Eine zentrale Herausforderung für langfristig angelegte Forschungsprojekte, wie sie im Akademienprogramm der Bund-Länder-Kommission in Deutschland mit Laufzeiten zwischen 15 und 25 Jahren üblich sind, ist die Nachhaltigkeit von Systemarchitekturen in einer ständig fortschreitenden Entwicklung von Standardisierung und Versionierung. PANDORA begegnet dieser Herausforderung mit einem entkoppelten Aufbau auf der Grundlage von relativ unabhängigen voneinander

agierenden Systemkomponenten, die bei Bedarf einfach ausgetauscht werden können, ohne die Grundfunktionalität zu gefährden. Auf der Ebene der Viewer können verschiedene Entwicklungen wie z.B. mirador(16) eingesetzt werden, ohne dass eine spezielle Anpassung notwendig ist. PANDORA setzt in Hinblick auf die langfristige Verfügbarkeit auf Standards aus dem Bereich des Semantik Web, die sich inzwischen weltweit durchgesetzt haben und damit sehr wahrscheinlich auch in Zukunft eine aktive Weiterentwicklung des Frameworks erlauben. Darüber hinaus ermöglichen diese Standards eine effiziente Vernetzung mit anderen Ressourcen im Web.

In dem Workshop sollen die einzelnen Komponenten des PANDORA Frameworks vorgestellt und deren Installation und Konfiguration erklärt werden. In einer Testumgebung, die für die Teilnehmer auf einem Server im Internet zur Verfügung stehen wird, können Beispieldatensätze gespeichert und die Funktionalität des Frameworks erprobt werden. Ebenfalls ist vorgesehen, die vorgestellte Architektur der Software intensiv zu diskutieren und mit anderen Lösungen für digitale Repositorien/Präsentationsumgebungen zu vergleichen.

Für die gewinnbringende Teilnahme sind Grundkenntnisse in Semantik Web Technologien sowie Kenntnisse der verwendeten Standards und/oder Open Source Software von Vorteil. Der Workshop eignet sich für eine Gruppe bis etwa 15 Personen. Die Teilnehmer sollten einen eignen Rechner/Laptop mit Verbindung zum Internet zur Verfügung haben, um im interaktiven Teil des Workshops die Funktionalitäten von PANDORA selber ausprobieren zu können. Die lokale Installation von zusätzlicher Software wird voraussichtlich nicht notwendig sein. Wichtige Informationen über die PANDORA Architektur können auch schon vorab in einem Video angesehen werden. (17)

Workshop Programm

Time	Title	Notes
9:00-9:30	Einführung	Gegenseitige Vorstellung. Einführung in das PANDORA Framework und Überblick zum Workshopverlauf
9:30-10:05	Ziele und Anwendungsbeispiele	Zielbeschreibung für den Workshop und Vorstellung von Anwendungsbeispielen
10:05-10:30	Technologische Herausforderungen	Vertiefende Einführung in die technologische Architektur von PANDORA - IIIF Presentation API, IIIF Image API, Manifest Service
10:30-10:45	Pause	
10:45-11:15	Einführung in den Übungsteil	Es wird eine Einführung für die Übung in kleinen Gruppen mit den IIIF Image und Presentation API gegeben.
11:15-11:40	Übungsteil / hands on	Hands-on Übung in kleinen Gruppen / Individuell Konkret: - Ein Manifest mit dem Javascript Client generieren - Ein Manifest mit dem IIIF Client visualisieren
11:40-12:00	Erfahrungsaustausch und Diskussion	Berichte aus den Kleingruppen
12:00-12:15	Pause	
12:15-1:00	Abschlussdiskussion, Fragen	Offene Fragen und Diskussion.

Organisatoren des Workshops:

Christopher Hanna Johnson, MA.
Projekt “Johann Friedrich Blumenbach-online” der ADW Göttingen
Geiststraße 10
37073 Göttingen
christopher.johnson@uni-goettingen.de oder chjohnson39@gmail.com
http://github.com/blumenbach
Forschungsinteressen: Semantik Web Technologien, Digitale Editionen, Softwareentwicklung, Cultural Heritage Studies

---

Dr. Jörg Wettlaufer
Digitisation Coordinator / Researcher
Akademie der Wissenschaften zu Göttingen (ADWG)
Göttingen Centre for Digital Humanities (GCDH)
Papendiek 16
37073 Göttingen
Germany
Tel. +49 551 39 20477 | 39 5366
jwettla@gwdg.de / skype: joewett
Forschungsinteressen: Digitale Geschichtswissenschaft,

Semantik Web Technologien, Digitale Editionen

Linkliste

http://fedorarepository.org/
https://www.w3.org/TR/json-ld/
https://www.w3.org/TR/sparql11-query/
http://iiif.io/api/presentation/2.1/
http://www.openannotation.org/spec/core/core.html
http://iiif.io/api/image/2.1/
https://www.w3.org/TR/ldp/#ldpc
https://www.w3.org/TR/rdf11-concepts/#section-rdf-graph
https://json-ld.github.io/normalization/spec/
http://www.rfc-editor.org/rfc/rfc5785.txt
https://github.com/blumenbach/iiif-manifest-service
http://iiif.io/api/presentation/2.1/#annotation-list
http://iiif.io/api/presentation/2.1/#layer
http://redis.io/topics/pubsub
https://github.com/blumenbach/
http://github.com/IIIF/mirador
Für ein einführendes Video zur PANDORA Architektur siehe: https://youtu.be/TEqUkiO6tcA

HowTo build a your own »Digital Edition Web-App«

Kampkaspar, Dario
kampkaspar@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Andorfer, Peter
Peter.Andorfer@oeaw.ac.at
Österreichische Akademie der Wissenschaften – Austrian Centre for Digital Humanities, Wien, Österreich

Baumgarten, Marcus
baumgarten@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Steyer, Timo
steyer@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Motivation

Aufgrund zahlreicher Sommer-Schulen, Workshops, DH-Studiengänge und vielfältiger online-Tutorials ist die Kodierung eines Textes in XML nach dem de-facto-Standard TEI ein oft anzutreffender Projektbestandteil. Was jedoch häufig fehlt sind einstiegsfreundliche Anleitungen, Tutorials, HowTos zu dem sich an die Kodierung anschließenden Themenkomplex der Publikation einer digitalen Edition. Die Frage nach dem »Wohin?« der oftmals in langer und mühsamer Arbeit erstellten Editionen betrifft vor allem jene Forschende, welche nicht Teil eines größer angelegten Projektes sind oder auch sonst über keine allzu starke Anbindung an eine gut institutionalisierte Forschungsinfrastruktur verfügen. Zwar entwickeln zunehmend mehr Institutionen, vielfach in Verbindung mit konkreten Projekten, Kompetenzen, Workflows und (technische) Infrastrukturen zur Veröffentlichung Digitaler Editionen, aufgrund chronisch knapper Finanzierung können oftmals aber nur wenige und in erster Linie nur eigene/interne Projekte hinreichend betreut werden.

Gleichzeitig kann in vielen Digitalen Editionsprojekten eine sehr starre Arbeitsteilung zwischen so genannten FachwissenschlafterInnen und TechnikerInnen beobachtet werden. Obwohl es sicherlich nicht als Nachteil bewertet werden kann, wenn jeder das tut, wofür er ausgebildet wurde und was sie bzw. er demzufolge auch gut kann, so besteht in einem stark arbeitsteiligen Umfeld die Gefahr asymmetrischer Kompetenzverhältnisse und daraus resultierender Abhängigkeiten. Sei es durch unrealistische Wünsche seitens der Fachwissenschaft, die aufgrund mangelnder technischer Kenntnisse an die Technik herangetragen werden. Oder sei es die Verzögerung des Arbeitsfortschritts aufgrund schleppender Implementierung basaler Technologien oder von editorischer Seite dringend benötigter Funktionalitäten.

Der hier vorgeschlagene Workshop versucht, beide Problembereiche aufzugreifen, indem gemeinsam mit den Teilnehmern, welche vorzugsweise ihre eigenen XML/TEI Daten mitbringen, eine auf der XML-Datenbank eXist basierte Web-Applikation zur Publikation eigener Editionen entwickelt wird.

Die Applikation

Die Anforderungen für eine solche Applikation stehen in engem Zusammenhang mit der im Kontext dieses Workshops verwendeten Vorstellung über die Eigenschaften und über potentielle Verwendungszwecke einer Digitalen Edition. Zur Erläuterung: Unter dem Begriff »Digitale Edition« sollen ein kohärenter Text oder mehrere kohärente Texte verstanden werden, die mittels XML/TEI kodiert wurden und worin in der Regel verschiedene Entitäten wie z.B. Personen, Orte, Werke oder ähnliches erfasst, deren Form und Textgenese beschrieben und die um weiterführende Erläuterungen, Annotationen und Anmerkungen ergänzt wurden. Eine solche Digitale Edition wird vorwiegend im ›close reading‹ rezipiert mit dem Zweck, ein tieferes Verständnis über den Text, dessen Inhalt sowie dessen Kontext und Entstehung zu erhalten. Abgesehen von einer solchen eher traditionellen Auseinandersetzung mit einer Digitalen Edition verfügt diese aber auch über den Mehrwert, systematisch und vor allem maschinell gelesen werden zu können.

Eine ›Digital Edition Web-App‹ sollte ganz generell die kodierten Texte in einer möglichst benutzerfreundlichen Art und Weise präsentieren und den »technischen Unterbau« dem Benutzer nicht aufbürden, wohl aber die computergestützte Weiterverarbeitung der Texte jederzeit ermöglichen. Konkreter formuliert heißt das, dass eine solche Anwendung folgende Anforderungen zu erfüllen hat.

Einstiegsseite

NutzerInnen sollen auf einer zentralen Einstiegsseite einen möglichst vollständigen Überblick über den kompletten Umfang der Edition erhalten. Dies ist insbesondere dann von großer Bedeutung, wenn die Edition aus mehreren Editionseinheiten besteht, wie zum Beispiel im Falle eines Briefwechsels.

In der im Zuge des Workshops zu entwickelnden Applikation wird das in Form einer ListView gelöst, welche sämtliche XML/TEI Dokumente bzw. ausgewählte Informationen aus dem teiHeader in einer von den NutzerInnen such-, filter- und sortierbaren Ansicht präsentiert. Von diesem Inhaltsverzeichnis gelangen die NutzerInnen dann über Verlinkung zu den einzelnen Dokumenten.

Responsive Design

Da Digitale Editionen im www verfügbar sind, muss davon ausgegangen werden, dass diese generell in digitaler Form, sprich auf einem PC, Notebook, Tablet, eventuell auch auf einem Smartphone gelesen werden. Insofern gilt es, den kodierten Text in einer leserfreundlichen Darstellung anzuzeigen, die die verschiedenen Formate der Anzeigegeräte berücksichtigt (womit einige der Grundlagen des sog. ›responsive design‹ berücksichtigt werden müssen). Andererseits darf aber der Wunsch vieler Nutzer, die Inhalte »klassisch« auf Papier zu nutzen, nicht vergessen werden.

Die digitale Darstellung im Web eröffnet indes auch die Möglichkeit für dynamische, sprich von den Nutzer/innen frei konfigurierbare, Darstellungsweisen. Abhängig vom konkreten Mark-Up können, um nur ein paar Beispiele zu nennen, etwa Anmerkungen ein- oder ausgeblendet, Abkürzungen aufgelöst, oder Korrekturschritte ausgeblendet werden.

In der ›Digital Edition Web-App‹ wird mittels XSLT Transformation aus den XML Dateien eine HTML Dokument ›on the fly‹ generiert. Diese ›DetailView‹ verfügt, sofern aufgrund des Markups des Ausgangsdokumentes möglich, über ein Navigationsmenü, welches eine rasche Orientierung im Text ermöglicht. Über ein weiteres Menü können außerdem verschiedene Darstellungsoptionen (de)aktiviert werden.

Suche

Die Möglichkeit, eine digitale Edition in ihrer Gesamtheit im Volltext durchsuchen zu können, wird häufig als einer der größten Vorzüge einer digitalen Edition beschrieben. Zusätzlich zu einer so genannten »einfachen Suche« wird darüber hinaus auch gerne eine »erweiterte Suche« angeboten, welche eine spezifizierte Suche wie zum Beispiel nur in Anmerkungen oder über Metadaten ermöglicht.

Aufgrund der Integration der Volltext-Suchengine Lucene in die Datenbanksoftware eXist-db ist die Realisierung sowohl einer »einfachen« wie auch einer »erweiterten« Suche im Rahmen der ›Digital Edition Web-App‹ einfach zu bewerkstelligen, wobei die Spezifika der »erweiterten« Suche vom konkreten Markup der einzelnen Editionen abhängt.

Einige grundlegende Überlegungen zum Erstellen einer Suche werden hierbei anhand konkreter Beispiele mit den Teilnehmern diskutiert und demonstriert werden.

Register

Neben einer Volltextsuche bieten viele digitale Editionen auch eine registerbasierte Suche an, mit deren Hilfe etwa gezielt Personen oder Orte in der Edition identifiziert werden können.

Je nach Art der Daten wird ein solches Register auf verschiedene Weisen demonstriert werden.

PDF-Erzeugung

Als Nachteil einer digitalen Edition wird oft angesehen, dass ihr die Möglichkeit, einfache Anmerkungen – ähnlich einem eigenen Studienexemplar – anzubringen, fehlt. Aus diesem und anderen Gründen wird häufig die HTML-Seite ausgedruckt.

Im Rahmen des Workshops werden hierzu zwei verschiedene Lösungswege kurz umrissen, ohne jedoch weiter ins Detail gehen zu können: Einerseits handelt es sich um ein für den Druck spezifisch erarbeitetes CSS-Stylesheet (»print-CSS«), andererseits die Generierung einer Datei für das Satzprogramm LaTeX.

Schnittstellen

Da die Texte in einer (einigermaßen) standardisierten Art und Weise kodiert sind, können diese auch maschinell prozessiert werden. Dafür ist es notwendig, dass nicht nur eine HTML Darstellung der Daten veröffentlicht wird, sondern auch die eigentlichen XML/TEI-Daten.

Die ›Digital Edition Web-App‹ wird ihre Daten über die in der eXist-db integrierte ›REST-Style Web API‹ veröffentlichen.

Ziel und Zielgruppe des Workshops

Ziel des Workshops ist es, den TeilnehmerInnen einen ersten Einblick in weit verbreitete Workflows, Technologien und Terminologien sowie Konzepte zur Umsetzung der genannten Funktionalitäten zu vermitteln. Sie erhalten somit Grundlagen zur Weiterentwicklung oder auch Beurteilung anderer Plattformen und Tools.

Die von den TeilnehmerInnen im Zuge des Workshops erarbeitete Web-App wird – auch aufgrund der Heterogenität der von den TeilnehmerInnen gestellten Daten – keine produktionsreife Applikation sein, die alle Aspekte einer digitalen Edition umsetzt. Allerdings bildet die im Workshop teilweise selbst geschriebene Software eine solide Basis für weiteres Selbststudium, woraus sich später für die einzelnen Teilnehmer oder Institutionen einfache, aber auf die spezifischen Bedürfnisse zugeschnittene Plattformen entwickeln können.

Die TeilnehmerInnen des Workshops sollten über Erfahrung in der Kodierung in XML/TEI verfügen und im besten Fall an einem konkreten Projekt arbeiten und über XML/TEI Dateien verfügen, auf deren Grundlage sie im Workshop ihre eigene ›Digital Edition Web-App‹ entwickeln können.

Ablauf und Teilnehmeranzahl

Die TeilnehmerInnen erhalten vorab eine detaillierte Anleitung zur Installation der notwendige Software (eXist-db).

Im eigentlichen Workshop werden die jeweiligen Arbeitsschritte von einem der Organisatoren live vorgeführt (dafür wird ein Beamer benötigt). Die konkreten Inhalte orientieren sich dabei an dem gleichnamigen Blog (Andorfer/Kampkaspar 2016), welcher von den Organisatoren im Rahmen der TEI-Konferenz 2016 offiziell präsentiert wurde.

Während des Workshops werden wir bei auftretenden Fragen und Problemen den Teilnehmenden helfend zur Seite stehen. Um eine möglichst gute Betreuung der TeilnehmerInnen gewährleisten zu können, sollte die Teilnehmerzahl 25 nicht überschreiten.

Organsiatoren

Peter Andorfer

hat im Zuge seiner Dissertation eine digitale Edition erstellt und war im Editionsprojekt »Die Korrespondenz von Leo von Thun-Hohenstein« für die technische Umsetzung des Projektes (Entwicklung der Web-Applikation) verantwortlich. Gemeinsam mit Dario Kampkaspar schreibt er außerdem für den Blog »HowTo build a digital edition web app«.

Dario Kampkaspar

erstellt im Rahmen seines Dissertationsprojektes eine Edition einer frühneuzeitlichen Handschrift. An der HAB ist er im Rahmen zweier Projekte (Andreas Bodenstein von Karlstadt; Johannes Rist) intensiv mit Edition und Entwicklung beschäftigt. Gemeinsam mit

Peter Andorfer schreibt er außerdem für den Blog »HowTo build a digital edition web app«.

Marcus Baumgarten

ist langjähriger Mitarbeiter an der HAB und betreut unterschiedliche Editionsprojekte. Zur Zeit arbeitet er in einem Kooperationsprojekt mit dem historischen Seminar der Universität Freiburg (die »Tagebücher des Fürsten Christian II. von Anhalt-Bernburg«) und gemeinsam mit dem Leibniz-Institut für europäische Geschichte in Mainz (»Digitale Edition europäischer Religionsfrieden zwischen 1500 - 1800«).

Gemeinsam mit Timo Steyer und Studierenden der TU Braunschweig betreibt er das Weblog www.digital-ist-besser.net

Timo Steyer

ist aktuell in den Bereichen Metadaten und Datenmodellierung im Forschungsverbund Marbach Weimar Wolfenbüttel am Standort Wolfenbüttel tätig. In diesem Kontext beschäftigt er sich mit Fragen und Methoden zu den Themen der Interoperabilität von digitalen Editionen und der Retrodigitalisierung von bereits im Druck vorliegenden Editionen (z. B. »Controversia et Confessio« und »Die Briefe der Fruchtbringenden Gesellschaft«).

Bibliographie

Andorfer, Peter / Kampkaspar, Dario (2016): *How to build a Digital Edition Web-App* http://www.digital-archiv.at/howto-build-a-digital-edition-web-app/ .

Nachhaltiges Management von Bildmetadaten mit XMP, exiftool und Fotostation

Pohl, Oliver
opohl@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Schrade, Torsten
torsten.schrade@adwmainz.de
Akademie der Wissenschaften und Literatur Mainz, Deutschland

Das Corpus Vitrearum Deutschland als Use Case

Das interakademische Vorhaben zur mittelalterlichen und frühneuzeitlichen Glasmalereiforschung „Corpus Vitrearum Medii Aevi“ (CVMA) in den Arbeitsstellen Potsdam und Freiburg steht derzeit der Herausforderung gegenüber, die im neu eingerichteten CVMA-Online-Bildarchiv [1] hinterlegten Bilddateien samt Metadaten für die Langzeitarchivierung und -lesbarkeit vorzubereiten.

Das Hauptziel des CVMA ist es, alle Glasmalereien des Mittelalters fotografisch zu erfassen, zu dokumentieren und zu edieren, um nicht nur das kulturelle Erbe zu bewahren und ins digitale Zeitalter zu überführen, sondern auch räumliche Distanzen zu überbrücken und die Glasmalereifotografien samt ihrer Dokumentation der Öffentlichkeit zugänglich zu machen.

Für die zielgerichtete Dokumentation hat das CVMA ein eigenes Metadatenschema verfasst [2], welches auf etablierte Schemata wie Dublin Core (Weibel et al. 1998) und IPTC (IPTC 2014) aus der professionellen Fotografie aufbaut und diese mit einem speziell auf die wissenschaftlichen Bedarfe der Glasmalerei ausgerichteten eigenen Namensraum erweitert. Dabei werden die Metadaten in das zugehörige Bild integriert, um so zusätzliche Abhängigkeiten von Datenbank- oder anderer Verwaltungssoftware zu vermeiden. Dadurch wird die plattformabhängige Nutzung und maschinelle Interpretierbarkeit des Datenbestands gefördert, was nötig ist, damit die Daten auch in Zukunft von Mensch und Maschine nachgenutzt werden können (Library of Congress und National Science Foundation 2003).

XMP

Zur Lösung der Herausforderungen im Bereich Langzeitarchivierung orientiert sich das CVMA an den bereits bestehenden Lösungen aus der digitalen Dokument-Langzeitarchivierung von PDF-Dateien (Braun et al. 2010; ISO 2011), bei welcher sämtliche genutzte Abbildungen, Schriftarten und Metadaten mit in das Dokument eingebettet werden, sodass diese Datei letztendlich alle Inhalte originalgetreu ohne Zuhilfenahme dritter Anwendungen anzeigen kann.

Der PDF-Standard (Adobe Systems 2006) nutzt für die Integration von Metadaten die eXtensible Metadata Platform (XMP), mit welcher die zum Dokument zugehörigen Metadaten als RDF/XML kodiert in das Dokument eingebettet werden (Adobe Systems 2005; Bright 2006) ohne die digitale Interpretierbarkeit der zu beschreibenden Datei zu beeinflussen. Wird die Datei von einem Programm geöffnet, welches den XMP-Standard nicht unterstützt, wird diese Datei dennoch angezeigt. Umgekehrt ist es möglich, dieselbe Datei mit einem Texteditor zu öffnen, um die XMP Daten in Reintextform anzuzeigen.

Die Nutzung von XMP hat sich bereits auf mehreren Ebenen etabliert. Zu einem werden beim Abspeichern von Digitalfotografien EXIF- und IPTC-Daten von Digitalkameras als XMP in die erzeugten Bilddateien geschrieben (Tesic 2005) und können so durch Bildverwaltungssoftware wie Adobe Bridge [3] oder FotoWare Fotostation [4] gelesen und verwaltet werden. Des Weiteren wird XMP in der Digital Asset Management Community für die einfache Verwaltung von Dateibeständen bevorzugt genutzt (Regli 2009). Durch die Einheit von Dokument und Metadaten verringert sich der Aufwand bei einem Datentransfer sämtlicher Abhängigkeiten auf das Kopieren des Datenbestands (Binder 2006) und hält gleichzeitig die Möglichkeit offen, Datenbanken aus den vorliegenden XMP-Daten zu erstellen und auf die XMP-Daten aufbauende Tools zu implementieren (Abdillah 2013). Der Kern von XMP liegt inzwischen auch als ISO-Standard vor. [5] Das CVMA-Onlinebildarchiv dient hierfür als Beispiel, da sich die relationale Datenbank der Online-Plattform aus den XMP-Daten des verfügbaren Bildbestandes speist.

Ein weiterer Faktor für die internationale Verbreitung des XMP-Standards ist seine Erweiterbarkeit. Demnach ist es möglich, eigene Metadatenschemata als XMP an Dateien anzuhängen. Zwar ist dieser Anhang als RDF/XML kodiert, allerdings besteht die Restriktion kein RDFS und OWL nutzen zu können (Eriksson 2007). Das einzig gültige Subjekt in den RDF-Tripeln in XMP ist die zu beschreibende Ressource, also die Datei selbst. Es ist also nicht möglich Ontologien wie im herkömmlichen Semantic-Web-Kontext zu verwenden. Für das CVMA-Bildarchiv genügt es jedoch, die beschreibenden Metadaten für die einzelnen Bildressourcen anzulegen und so den Funktionsumfang des XMP-Standards voll auszuschöpfen.

XMP Workflow beim CVMA

In Deutschland haben die zwei CVMA-Arbeitsstellen in Potsdam und Freiburg unterschiedliche Workflows implementiert, um XMP-Daten in die Bilddateien einzupflegen. Beide benutzen zwar dasselbe Metadatenschema, allerdings unterscheiden sich die Tools zur Eingabe der Metadaten und die damit verbundenen Arbeitsabläufe.

Um sich für das Vorhaben der deutschen CVMA-Arbeitsstellen zu eignen, muss eine entsprechende Metadatenbearbeitungssoftware folgende Kriterien erfüllen: a) Die Software muss fähig sein, XMP-Metadaten zu schreiben und auszulesen. b) Die Software muss die Anzeige und Metadatenmanipulation von gängigen Bilddateiformaten wie TIFF und JPG unterstützen. c) Die Software muss die Möglichkeit zur Konfiguration eines eigenen Metadatenschemas anbieten. d) Optional: Die Software bietet die Möglichkeit zur lokalen Recherche mit dem im c) angelegten Metadatenschema.

Das Team in Freiburg nutzt die Freeware exiftoolGUI [6], welche eine konfigurierbare graphische Oberfläche für das Kommandozeilentool exiftool [7] bietet. exiftool selbst ist ein Programm zum Auslesen und zur Manipulation von Bildmetadaten und unterstützt das Lesen und Schreiben von XMP-Daten. Über eine Perl-Konfigurationsdatei kann das anzuwendende Metadatenschema für Datenmanipulationen via exiftool und somit auch exiftoolGUI festgelegt werden. Die inhärenten Vorteile von exiftool und exiftoolGUI sind deren Offenheit, Konfigurierbarkeit und Plattformunabhängigkeit. Die Nutzung von exiftoolGUI für größere Bildbestände ist jedoch eher unkomfortabel, da dieses Tool ausschließlich zur Metadatenmanipulation und nicht als Bildverwaltungssoftware ausgelegt ist.

Die CVMA-Arbeitsstelle in Potsdam nutzt hingegen die proprietäre Software FotoStation von FotoWare (Version 8.0). Zwar liegt der Quellcode dieser Software nicht offen, jedoch besteht auch bei diesem Tool die Möglichkeit zur Konfiguration eigener Metadatenschemata. FotoStation bietet deutlich mehr Bedienkomfort im Vergleich zu exiftoolGUI und kann als Bildverwaltungssoftware genutzt werden, welche den lokalen Bildbestand samt Metadaten für Recherchen indexiert und die Bilder auf einem digitalen Lichttisch darstellt. Für die Konfiguration des Metadatenschemas bietet FotoStation graphische Editoren, mit welchem

nicht nur die zu verwendenden Namensräume und Felder definiert werden können, sondern auch das Metadateneingabeinterface frei gestaltet werden kann. Dabei ist es möglich kontrollierte Vokabulare zu anzulegen und die Nutzereingaben über reguläre Ausdrücke validieren zu lassen. Weiterhin lässt sich die Gesamtkonfiguration von Fotostation leicht ex- und importieren, sodass sämtliche Informationen zum verwendeten Metadatenschema, Editorinterface, Vorschlagslisten, Bearbeitungsaktionen und Rerchercheeinstiege leicht innerhalb des Teams ausgetauscht und aktualisiert werden können.

Ziele des Workshops

Die Teilnehmer an diesem Workshop sollen einen Einstieg in die nachhaltige Bildmetadatenverwaltung mit XMP erhalten. Als Übung wird mit den Teilnehmern gemeinsam zuerst ein Beispiel-Metadatenschema definiert. Anschließend werden die Teilnehmer in die Benutzung von exiftoolGUI und FotoStation eingeführt, um daraufhin das zuvor definierte Metadatenschema für beide Tools umzusetzen und zu testen. Weiterhin wird das Abfragen von XMP-Metadaten mit dem exiftool über die Kommandozeile geübt, um so das Potential von XMP für die Erstellung von Tools, Services oder Datenbanken für Bildbestände aufzuzeigen.

Das CVMA-Deutschland erhofft sich, Interessierten und ähnlichen Projekten den Einstieg in die Erstellung und Manipulation von XMP-Daten zu erleichtern und gleichzeitig weitere Lösungsansätze und Tools für den Umgang mit XMP-Daten kennenzulernen.

Für die Teilnahme an diesem Workshop werden lediglich Grundkenntnisse in XML vorausgesetzt. Die Kenntnisse einer Programmier- oder Skriptsprache sind zwar von Vorteil, aber nicht erforderlich. Den Teilnehmern entstehen für die Nutzung von Software während des Workshops keine Kosten, da exiftool und exiftool GUI kostenfrei sind und Fotostation für 14 Tage kostenfrei getestet werden kann. Der Workshop ist für maximal 20 Teilnehmer ausgelegt.

Forschungsinteressen

Oliver Pohl ist wissenschaftlicher Mitarbeiter bei TELOTA an der Berlin-Brandenburgischen Akademie der Wissenschaften und betreut dort das CVMA, das Langzeitvorhaben Corpus Coranicum sowie das Kooperationsprojekt Paleocoran mit dem Collège de France. Seine Forschungsinteressen sind Webtechnologien und Semantic Web Technologien für digitale Geisteswissenschaften als auch maschinelle Übersetzung.

Torsten Schrade ist Leiter der Digitalen Akademie der Mainzer Akademie der Wissenschaften und der Literatur und beschäftigt sich vorrangig mit dem Forschungsdatenmanagement und dem Einsatz von Webtechnologien für die geisteswissenschaftliche Grundlagenforschung. Daneben zählen Methoden und Programmierparadigmen der agilen Softwareentwicklung sowie die Technologien des Semantic Web zu seinen Forschungsinteressen.

Fußnoten

1. http://www.corpusvitrearum.de
2. http://www.corpusvitrearum.de/cvma/1.1/ (Stand: 25.08.2016)
3. http://www.adobe.com/de/products/bridge.html
4. http://www.fotoware.com/products/fotostation-client
5. http://www.iso.org/iso/catalogue_detail?csnumber=57421
6. http://u88.n24.queensu.ca/~bogdan/ bzw. https://hvdwolf.github.io/pyExifToolGUI/
7. http://www.sno.phy.queensu.ca/~phil/exiftool/

Bibliographie

Abdillah, Leon Andretti (2013): „PDF Articles Metadata Harvester“, in: *arXiv Preprint arXiv:1301.6591*. http://arxiv.org/abs/1301.6591 [letzter Zugriff 25. August 2016].

Adobe Systems (2005): *Extensible Metadata Platform (XMP) Specification*. Adobe Systems https://partners.adobe.com/public/developer/en/xmp/sdk/XMPspecification.pdf [letzter Zugriff 25. August 2016].

Adobe Systems (2006): *PDF Reference, Sixth Edition: Adobe Portable Document Format Version 1.7* http://wwwimages.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/pdf_reference_1-7.pdf [letzter Zugriff 25. August 2016].

Binder, Jennifer (2006): „Exchanging Assets and Metadata across Platforms“, in: *Journal of Digital Asset Management* 2 (5): 215–18 10.1057/palgrave.dam.3650045.

Braun, Kim / Buddenbohm, Stefan / Dobratz, Susanne / Herb, Ulrich / Müller, Uwe / Pampel, Heinz / Schmidt, Birgit

(2010): *DINI-Zertifikat Dokumenten-Und Publikationsservice 2010* https://pub.uni-bielefeld.de/publication/2491543 [letzter Zugriff 25. August 2016].

Bright, Jason (2006): „First Steps: XMP“, in: *Journal of Digital Asset Management* 2 (3–4): 198–202 10.1057/palgrave.dam.3650025.

Eriksson, Henrik (2007): „The Semantic-Document Approach to Combining Documents and Ontologies“, in: *International Journal of Human-Computer Studies* 65 (7): 624–639.

IPTC (2014): *IPTC - NAA Information Interchange Model Version 4.2*. International Press Telecommunicatoins Council http://www.iptc.org/std/IIM/4.2/specification/IIMV4.2.pdf [letzter Zugriff 25. August 2016].

ISO (2011): *ISO 19005-1:2005 - Document Management -- Electronic Document File Format for Long-Term Preservation -- Part 1: Use of PDF 1.4 (PDF/A-1)* http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=38920 [letzter Zugriff 25. August 2016].

Library of Congress / National Science Foundation (2003): *It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation* http://www.digitalpreservation.gov/documents/about_time2003.pdf [letzter Zugriff 25. August 2016].

Regli, Theresa (2009): „The State of Digital Asset Management: An Executive Summary of CMS Watch's Digital Asset Management Report“, in: *Journal of Digital Asset Management* 5 (1): 21–26.

Tesic, Jelena (2005): „Metadata Practices for Consumer Photos“, in: *IEEE MultiMedia* 12 (3): 86–92.

Weibel, Stuart / Kunze, John / Lagoze, Carl / Wolf, Misha (1998): *Dublin Core Metadata for Resource Discovery* http://www.rfc-editor.org/info/rfc2413 [letzter Zugriff 25. August 2016].

open your data, open your code: Offene Lizenzierung für geisteswissenschaftliche Projekte

Hannesschläger, Vanessa
vanessa.hannesschlaeger@oeaw.ac.at
ACDH-OEAW Austrian Centre for Digital Humanities, Österreichische Akademie der Wissenschaften, Österreich

Losehand, Joachim
joachim@losehand.at
creative commons Austria

Kamocki, Paweł
kamocki@ids-mannheim.de
IDS Mannheim, Deutschland

Scholger, Walter
walter.scholger@uni-graz.at
ZIM-ACDH Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Karl-Franzens-Universität Graz, Österreich

Witt, Andreas
witt@ids-mannheim.de
IDS Mannheim, Deutschland

Amini, Seyavash
amini@ivocat.de
e-Infrastructures Austria

Einleitung

In Diskussionen über digitale Nachhaltigkeit sind rechtliche Fragen von zunehmender Bedeutung. Häufig entscheiden die rechtlichen Rahmenbedingungen, ob, wie und wie lange Daten und Programme, die im Rahmen von digitalen geisteswissenschaftlichen Forschungsprojekten entwickelt werden, verfügbar sind. Gerade im digitalen Raum, der nationale Grenzen überschreitet, ergibt sich - auch aufgrund der territorialen Beschränkung des jeweiligen Urheberrechts - die Notwendigkeit

neuer rechtlicher Regelungen, die einerseits das Urheberrecht des Forschenden schützen, andererseits die Wiederverwendbarkeit ihrer Arbeiten sicherstellen sollen. Offene Lizenzierungsmodelle bieten hier Lösungen, die internationale Forschung an lokal erzeugten Daten und mit individuell entwickelten Softwares ermöglichen.

In den Geisteswissenschaften, in denen die Publikation von und die Arbeit mit "Rohdaten" im Rahmen der digitalen Wende substantiell an Bedeutung gewonnen haben, beginnt die Forschungscommunity zunehmend, sich diesem Thema im Kontext von Open Access-Diskussionen zu widmen. Obwohl das Bewusstsein für die Notwendigkeit der Auseinandersetzung mit diesen rechtlichen Aspekten zunimmt, fehlt den Forschenden oft der Überblick über die unterschiedlichen Möglichkeiten der Lizenzierung ihrer Daten. Creative Commons-Lizenzen sind weltweit das etablierteste Modell. Während sie in vielen Fällen eine gute Lösung sind, kommen sie häufig auch nur deshalb zum Einsatz, weil den Forschenden Informationen über andere Lizenzierungsmodelle fehlen. Auch wenn Creative Commons-Lizenzen angemessen sind, herrscht in vielen Fällen oft Unklarheit und Unsicherheit über die Wahl der konkret geeigneten Lizenz. Noch komplexer ist die Landschaft an verfügbaren Software-Lizenzen.

Dieser Workshop möchte zur Information und Aufklärung der Community beitragen, indem zuerst ein allgemeiner Überblick über die unterschiedlichen nationalen rechtlichen Rahmenbedingungen und über verfügbare Daten- und Software-Lizenzen gegeben wird, bevor im zweiten Teil lizenzierte Beispielprojekte vorgestellt werden. Eingegangen wird auch auf die unterschiedlichen Materialtypen (z.B. Manuskript-, Photodigitalisate) und den umgang mit "verwaisten Werken". Erwartet wird ein Publikum, das sich einen Überblick über die Lizenz-Möglichkeiten, die sich für einzelne Projekte bieten, verschaffen und die passende Lizenz für das jeweils eigene Projekt finden möchte. Dazu dient der dritte Abschnitt des Workshops, der in der Art einer offenen Sitzung gestaltet ist, in der nach einer Einführung zu Lizenzierungstools Projekte aus dem Publikum diskutiert werden. Die Beitragenden sind Expert/innen für Lizenzierungs- und Rechtsfragen und wünschen sich den Austausch mit Forschenden, die sich ebenfalls mit diesen Themen beschäftigen.

Daten-Lizenzen

Creative Commons (CC)

Creative Commons (CC) ist eine Non-Profit-Organisation, die eine Auswahl an Standard-Lizenzverträgen zur öffentlichen Nutzung von Werken für juristische Laien entwickelt hat. Der Ausgangspunkt dafür war, dass "[t]he idea of universal access to research, education, and culture is made possible by the Internet, but our legal and social systems don't always allow that idea to be realized." (https://creativecommons.org/about/) CC-Lizenzen sind standardisierte Lizenzen, die durch die Creative-Commons-Stiftung entwickelt und ausformuliert wurden und Rechteinhabenden zur freien Verwendung zur Verfügung stehen. Sie sind vorrangig für den Einsatz bei digitalen Werken und der Verbreitung im Internet geschaffen worden, lizenzieren aber gleichzeitig auch die Verwendung im nicht-digitalen Bereich (Druckwerke); sie umfassen alle (wesentlichen) urheberrechtlichen und leistungsschutzrechtlichen Aspekte.

CC-Lizenzen der Versionen 1.0 bis 3.0 wurden vielfach mit nationalen Recht akkordiert (d.h. "portiert") und so in spezifischen nationalen Varianten zur Verfügung gestellt. Ab Einführung der aktuellen Version 4.0 erfolgten keine spezifischen nationalen Adaptierungen mehr, um eine einheitliche, international gültige Lizenzierung und Rechtssicherheit zu gewährleisten. In jedem Fall sind aber - wie bei allen Einzel- und Massenlizenzen - die jeweils geltenden nationalen urheberrechtlichen bzw. internationalen Regelungen vorrangig und deshalb zu beachten.

Alle nach geltendem Recht urheberrechtlich schutzwürdigen Werke können durch die Rechteinhabenden zu jeder Zeit mit Creative Commons (neu) lizenziert werden. Eine einmal erteilte CC-Lizenz kann grundsätzlich nicht widerrufen und damit erteilte Nutzungsrechte können grundsätzlich nicht eingeschränkt werden, jedoch kann die Lizenz in eine weniger einschränkende Lizenz umgewandelt werden. Davon sollte jedoch nur in einzelnen Ausnahmefällen Gebrauch gemacht werden.

CC-Lizenzen bestehen aus mehreren Modulen, wobei die Zusammensetzung nicht von den Lizenzgebenden gewählt werden kann, sondern vorgegeben ist.

In den Digital Humanities hat es sich zum de-facto-Standard entwickelt, Projekte, Daten und Ergebnisse unter CC-Lizenzen zur

Verfügung zu stellen, wobei die Unterschiede zwischen den verschiedenen Versionen, portierten und nicht portierten Fassungen und unterschiedlichen Lizenz-Inhalten oft zu Unklarheiten führen. In diesem Workshop werden daher die Grundlagen des CC-Modells erklärt und zahlreiche weitergehende Aspekte erläutert.

Digital Peer Publishing (DiPP/ DPPL)

Das Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz) stellt zur Erleichterung der Publikation von Open Access-Journals das System Digital Peer Publishing (DiPP) zur Verfügung, in dessen Rahmen auch die Digital Peer Publishing Lizenzen (DPPL) entwickelt wurden. Aktuell ist die Version 3.0. Die Grundlage dieser Lizenzen bildet das deutsche Recht, was - anders als CC - im internationalen Bereich häufiger zu Problemen führen kann. Um den Teilnehmenden eine breitere Perspektive zu ermöglichen, werden als eine mögliche Alternative zu Creative Commons die DPP-Lizenzen vorgestellt.

Software-Lizenzen

Es ist zu beachten, dass die oben beschriebenen Datenlizenzen nicht zur Lizenzierung von Software geeignet sind, da sie diverse Software-spezifische Aspekte nicht berücksichtigen. Es müssen für diesen Zweck daher spezifische Software-Lizenzen verwendet werden, die in einem weiteren Teil des Workshops vorgestellt werden.

Freie Lizenzen haben im Bereich der Software bereits eine längere Tradition, als dies bei Daten und anderen Inhalten der Fall ist: Die ersten öffentlichen Software-Lizenzen gab es in den 1980er Jahren. Heute ist eine Fülle an freien Softwarelizenzen in Gebrauch. Nach gegenwärtigen Standards gilt eine Lizenz dann als Open Source, wenn sie den von der Open Source Initiative entwickelten Kriterien entspricht (https://opensource.org/osd-annotated). Die Wichtigsten davon sind: Die Möglichkeiten des Zugangs, der Verbreitung und der Adaptierung von Quellcode.

Des Weiteren können Open Source Lizenzen in drei Kategorien eingeteilt werden:

- Freizügig (permissive): erlaubt breite Verwendung der lizenzierten Software (z.B. BSD, MIT oder Apache Lizenzen)
- Starkes Copyleft: “virale” Lizenzen, die den Benutzenden auferlegt, modifizierten Code unter einer kompatiblen Copyleft-Lizenz zu veröffentlichen (GNU GPLs sind die wichtigsten Lizenzen dieser Gruppe)
- Schwaches Copyleft: Lizenzen, die die Benutzenden verpflichten, modifizierte Software unter einer kompatiblen Copyleft-Lizenz zu veröffentlichen, aber die Verbindung mit Bibliotheken erlauben, die andere Lizenzen verwenden (z.B. GNU LGPL)

Aufgrund dieser breiten Auswahl an Möglichkeiten, von denen sich einige nur durch Formalitäten unterscheiden, variiert die Lizenzierungspraxis zwischen Communities und Projekten stark. Das ist insofern problematisch, als dadurch in großen, verteilten Projekten oft komplexe Kompatibilitätsprobleme entstehen. Es ist daher notwendig, gemeinsame “best practices” zu entwickeln, um die Bestrebungen innerhalb der Digital Humanities Community zu harmonisieren.

Lizenzierungstools

Um den Teilnehmenden im praktischen Teil des Workshops nicht nur konkrete Beratung für aktuelle Projekte anbieten zu können, sondern ihnen auch Unterstützung für die Entscheidung von Lizenzierungsfragen im Rahmen zukünftiger Projekte und Kontexte zur Verfügung zu stellen, werden ausgewählte Lizenzierungstools präsentiert und unmittelbar unter Anleitung der Beitragenden getestet. Zu den wesentlichsten Hilfsmitteln gehören:

- Europeana Available Rights Statement < http://pro.europeana.eu/share-your-data/rights-statement-guidelines >
- CLARIN License Category Calculator < https://www.clarin.eu/content/license-categories >
- Licentia License Tool < http://licentia.inria.fr/ >
- ELRA License Wizard < http://wizard.elda.org/ >
- Public License Selector < https://ufal.github.io/public-license-selector/ >

Ablauf

10.00 Uhr: Einleitung

10.15 Uhr: Vorstellung der Lizenzmodelle

- Lizenzen für geisteswissenschaftliche Inhalte (Joachim Losehand/Seyavash Amini)
- Free/Open Source Software Licensing (Paweł Kamocki)
- Diskussion

11.45 Uhr: Pause
13.00 Uhr: Beispielprojekte

- Lizenzierungsmodelle in CLARIN (Andreas Witt)
- Das Geisteswissenschaftliche Asset Management-System GAMS: Objekte, Digitalisate und Vervielfältigung (Walter Scholger)
- Internationale Werke: Lizenzierung des Abstractbands der TEI-Konferenz 2016 (Vanessa Hannesschläger)
- Diskussion

14.30 Uhr: Pause
15.00 Uhr: hands-on: select your license

- Einführung: Lizenzierungstools (Walter Scholger)
- Offene Diskussion & Beratungsrunde: Konkrete Beispiele aus dem Publikum

17.00 Uhr: Ende

Benötigte Infrastruktur, Teilnehmende

Erwartet wird ein großes Publikum, da der Workshop konkrete Beratung für individuelle Projekte vorsieht. Um jedoch einen lebendigen Austausch garantieren zu können, möchten wir die Zahl der Teilnehmenden auf 40 Personen beschränken.

Teilnehmende werden gebeten, ihre eigenen Laptops mitzubringen.

Benötigt wird ein Raum für 40 Personen mit WLAN und einem Beamer; ein zusätzlicher Computer mit Bildschirm & installiertem Skype (für die Zuschaltung von Seyavash Amini) wäre wünschenswert.

Eine Mittagspause von 11.45-13.00 Uhr und eine Kaffeepause von 14.30-15.00 Uhr ist geplant. Sollte für die Mittagspause keine Verpflegung verfügbar sein, werden die Teilnehmenden gebeten, sich selbst zu versorgen. Kaffee und Getränke in der Kaffeepause wären wünschenswert.

Beitragende

Vanessa Hannesschläger < vanessa.hannesschlaeger@oeaw.ac.at>
studierte Germanistik in Wien. Sie ist wissenschaftliche Mitarbeiterin des Austrian Centre for Digital Humanities der Österreichischen Akademie der Wissenschaften (ACDH-ÖAW) und dort für Rechts- und Lizenzierungsfragen zuständig. Schon im Rahmen eines vorangegangenen Forschungsprojekts am Literaturarchiv der Österreichischen Nationalbibliothek beschäftigte sie sich mit praktischen Fragen des Urheberrechts. Mitarbeit in CLARIN (CLARIN PLUS) und DARIAH (WG Thesaurus Maintenance). Zu ihren Forschungsinteressen gehören digitales Edieren, Text- und Datenmodellierung, das Archiv im digitalen Kontext, Vermittlungsstrategien in den DH sowie digitale Infrastrukturen.

Joachim Losehand < joachim@losehand.at>
ist Kulturhistoriker und studierte u.a. Klassische Archäologie, Alte Geschichte und Altertumswissenschaften in Tübingen, München und Wien. Zwischen 2003 und 2006 war er wissenschaftlicher Lektor und Redakteur, seit 2006 ist er Lehrbeauftragter und Lektor an Universitäten in Bremen, Oldenburg und Wien. 2009/10 war er Mitglied der Lenkungsgruppe im Aktionsbündnis „Urheberrecht für Bildung und Wissenschaft“, seit 2013 ist er Projektkoordinator und Referent für Urheberrecht u.a. im Verband Freier Radios Österreich (VFRÖ) sowie Projektleiter Science Commons bei creative commons Austria.

Paweł Kamocki < kamocki@ids-mannheim.de>
verfügt sowohl im Bereich des Rechts als auch im Bereich der Sprachwissenschaften über breites Fachwissen; derzeit ist er wissenschaftlicher Mitarbeiter am Institut für Deutsche Sprache in Mannheim und Lehr- und Forschungsassistent an der Descartes Universität in Paris und promoviert zu den rechtlichen Fragestellungen der Open Science. Er ist Mitglied des CLARIN Legal Issues Committee und arbeitete als rechtlicher Berater in zahlreichen anderen Projekten und Arbeitsgruppen (z.B. EUDAT, RDA, OpenMinTeD). Neben Urheberrecht und Datenschutz gilt sein Interesse auch den Sprachwissenschaften (insb. rechtliche Fachsprache).

Walter Scholger < walter.scholger@uni-graz.at>
studierte Geschichte und Angewandte Kulturwissenschaften in Graz und Maynooth

und ist administrativer Leiter des Zentrums für Informationsmodellierung - Austrian Centre for Digital Humanities an der Universität Graz. In Projekten, internationalen Workshops und universitärer Lehre widmet er sich rechtlichen Aspekten des digitalen Kulturerbes und Fragen offener digitaler Publikationsformen.

Er ist Mitglied in facheinschlägigen Arbeitsgruppen der Digital Humanities Dachverbände und internationaler Projekte (ADHO, DHd, ICARUS, DARIAH) zu rechtlichen Aspekten, digitalen Publikationen und Lehre im Bereich der Digital Humanities.

Andreas Witt < witt@ids-mannheim.de>

leitet den Programmbereich Forschungsinfrastrukturen am Institut für Deutsche Sprache in Mannheim und ist Honorarprofessor für Digital Humanities an der Universität Heidelberg. Seine Forschungsinteressen konzentrieren sich auf die Texttechnologie, die Digital Humanities, die Informationsmodellierung und Auszeichnungssprachen. Bei CLARIN-D ist er für die Arbeitsgruppe zu juristischen und ethischen Fragen beim Umgang mit digitalen Sprachressourcen verantwortlich.

Seyavash Amini < amini@ivocat.de>

ist Rechtsberater der Universitätsbibliothek Wien, Teamleiter des Clusters E - “Legal and Ethical Issues” im Projekt *e-Infrastructures Austria*, Berater der Geschäftsleitung einer Gruppe von Medienunternehmen in Hannover sowie Lehrbeauftragter an den Universitäten Wien und Hannover. Er beschäftigt sich mit Fragen des Informations-, Immaterialgüter-, Medien- und Datenschutzrechts. Im Rahmen der jüngsten Novelle des österreichischen Urheberrechts hat der Gesetzgeber einen von Seyavash Amini mitgestalteten Formulierungsvorschlag aufgegriffen und umgesetzt. Er wird sich per Skype zum Workshop zuschalten.

Bibliographie

Amini, Seyavash / Blechl, Guido / Losehand, Joachim (2015): *FAQs zu Creative-Commons-Lizenzen unter besonderer Berücksichtigung der Wissenschaft* https://phaidra.univie.ac.at/view/o:408042 [letzter Zugriff 25. August 2016].

Creative Commons. https://creativecommons.org/ [letzter Zugriff 25. August 2016].

DiPP - Digital Peer Publishing. http://www.dipp.nrw.de/ [letzter Zugriff 25. August 2016].

Kamocki, Paweł / Ketzan, Erik (2014): *Creative Commons and Language Resources: General Issues and What’s New in CC 4.0*. CLARIN Legal Issues Committee: White Paper Series http://clarin-d.de/images/legal/CLIC_white_paper_1.pdf [letzter Zugriff 25. August 2016].

Kamocki, Paweł / Ketzan, Erik / Witt, Andreas (2016): „Lizenzauswahlwerkzeuge für die digitalen Geisteswissenschaften“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 336–337 http://dhd2016.de/boa.pdf [letzter Zugriff 25. August 2016].

Klimpel, Paul (2013): *Free knowledge thanks to creative commons licenses: Why a non-commercial clause often won’t serve your needs*. Wikimedia Deutschland / iRights.info / CC DE. https://www.wikimedia.de/w/images.homepage/1/15/CC-NC_Leitfaden_2013_engl.pdf [letzter Zugriff 25. August 2016].

Klimpel, Paul / Weitzmann, John H. (2015): *Forschen in der digitalen Welt. Juristische Handreichung für die Geisteswissenschaften*. DARIAH-DE Working papers 12 https://irights.info/wp-content/uploads/2015/08/Forschen-in-der-digitalen-Welt-Juristische-Handreichung-Geisteswissenschaften-dwp-2015-12.pdf .

Panels

Aktuelle Herausforderungen der Digitalen Dramenanalyse

Willand, Marcus
marcus.willand@ilw.uni-stuttgart.de
Universität Stuttgart

Trilcke, Peer
trilcke@uni-potsdam.de
Universität Potsdam

Schöch, Christof
christof.schoech@uni-wuerzburg.de
Universität Würzburg

Rißler-Pipka, Nanette
nanette.rissler-pipka@ku.de
Katholische Universität Eichstätt-Ingolstadt

Reiter, Nils
nils.reiter@ims.uni-stuttgart.de
Universität Stuttgart

Fischer, Frank
ffischer@hse.ru
Higher School of Economics, Moskau

Zielstellung und Konzeption

Das hier vorgeschlagene Panel greift mit der Digitalen Dramenanalyse einen sich derzeit dynamisch entwickelnden Bereich der digitalen Literaturwissenschaften auf. Es setzt sich erstens zum Ziel, aktuelle Herausforderungen der Digitalen Dramenanalyse auf verschiedenen Ebenen vorzustellen, wobei insbesondere die Ebenen der dramatischen Gattung, der Netzwerkstrukturen und der dramatischen Figuren im Zentrum stehen werden. Zweitens möchte das Panel mit dem Publikum mögliche Lösungsansätze diskutieren, unter anderem durch Bezug auf vielfältige, vorhandene Erfahrungen mit der Analyse narrativer Texte. In der Summe wird das Panel einerseits eine Zwischenbilanz zum Stand der Forschung anbieten, andererseits auch im Sinne einer Konsolidierung des Forschungsfelds eine Agenda für die weitere Entwicklung formulieren, bei der es nicht zuletzt darum geht, Szenarien einer integrativen, mithin diverse methodische Ansätze synergetisch zusammenführenden Forschung, zu diskutieren.

Dazu wird das Panel eine Art Laborsituation fingieren, in der die Erkenntnisziele, Möglichkeiten und Grenzen unterschiedlicher methodischer Zugänge zu dem titelgebenden Forschungsbereich der digitalen Dramenanalyse zu Tage treten sollen: *Topic Modeling*, (soziale) Netzwerkanalyse und Analyse der Figurenrede. Diese Gegenüberstellung soll es dem Publikum erlauben, Grundannahmen und Perspektivierungen der jeweiligen Ansätze direkt zu identifizieren und in der Diskussion adressieren zu können. Welche Modellierung des dramatischen Textes liegt der Methode zugrunde? Welche Aspekte eines dramatischen Textes werden durch die jeweilige Methode eigentlich beobachtet, und welche nicht? Und welche Art von Aussagen macht sie möglich?

In den *digital literary studies* wird zwar häufig Methodenkritik geäußert, dies in der Regel aber nur mit Blick auf einzelne Methoden, auch wenn diese auf ganz unterschiedliche Forschungsgegenstände angewandt werden können. Dieses Vorgehen möchte das Panel invertieren, indem nicht eine Methode auf unterschiedliche Objekte, sondern unterschiedliche Methoden auf das gleiche Objekt angewandt werden: Digitalisierte Dramen zwischen 1700 und 1900 aus dem deutsch- und französischsprachigen Raum. Erreicht werden soll durch dieses Vorgehen eine systematische Aufarbeitung der Möglichkeiten einer methodisch reflektierten digitalen Dramenanalyse, die zugleich theoretische und methodologische Grundfragen der digitalen Analyse literarischer Texte im Allgemeinen thematisiert.

In drei Kurzvorstellungen sollen die folgenden Methoden von jeweils einer Forschergruppe des Panels vorgestellt werden, wobei zur besseren Vergleichbarkeit der drei methodisch unterschiedlich aufgestellten Arbeitsgruppen zeitlich und gattungsbezogen vergleichbare Textsammlungen analysiert werden. Zwar werden diese Verfahren jeweils anhand eines individuellen Teilkorpus vorgestellt, es ist jedoch zu berücksichtigen, dass sie alle auf der statistischen Analyse größerer Textmengen basieren.

Panelvorträge

Topic Modeling und Gattung (Christof Schöch, Nanette Rißler-Pipka)

Der Einsatz von *Topic Modeling* (Blei 2012) für die im weitesten Sinne inhaltliche Erschließung großer Sammlungen literarischer Texte zeigt zwei Dinge: Erstens sind die erzielten Ergebnisse, insbesondere die jeweils dominanten Typen der *Topics*, textsortenabhängig (Schöch 2016). So sind nicht-fiktionale expositorische Texte (bspw. Pressemitteilungen) durch abstrakte thematische *Topics* geprägt, fiktionale Erzähltexte (bspw. Romane) aber durch *Topics*, die sich auf narrative und deskriptive Motive beziehen. Auch dramatische Texte zeichnen sich durch ein eigenes Profil solcher Typen von *Topics* aus, in dem diskursive und metadiskursive *Topics* eine besondere Rolle spielen. Dieser Umstand schärft auch den Blick auf die spezifische, textuelle Funktionsweise der jeweiligen Gattung und literarischer Texte insgesamt.

Zweitens zeigt sich, dass sich einzelne dramatische Untergattungen wie Tragödie, Komödie oder Tragikomödie zwar in Bezug auf die jeweils dominanten Einzel *topics* unterscheiden (und beispielsweise jeweils ein unterschiedlich strukturiertes *Liebes-Topic* haben können). Zugleich fördert *Topic Modeling* aber keine scharfen Trennungslinien zu Tage, sondern zeigt auf, wie prototypisch gedachte Untergattungen in der Praxis unscharf ineinander übergehen können (Schöch, im Erscheinen). Beide genannten Phänomene sind bekannt, aber sowohl in methodischer bzw. informatischer als auch in literaturtheoretischer Perspektive derzeit nicht ausreichend klar erfasst und damit auch nicht empirisch überprüfbar.

Netzwerkanalyse (Frank Fischer, Peer Trilcke)

Die in den quantitativen Sozialwissenschaften entwickelten Verfahren der Netzwerkforschung zielen auf eine formale Analyse sozialer Strukturen (Wasserman / Faust 1994). Angewandt auf literarische Texte ermöglichen sie Strukturbeschreibungen, die aus einer signifikant anderen Perspektive erfolgen als traditionelle literaturwissenschaftliche Verfahren der semantikbasierten Strukturanalyse (z.B. Titzmann 1977), insofern sie nicht die semantische Organisation literarischer Texte, sondern die ästhetische Modellierung sozialer Formationen im Medium der Literatur analysieren (Trilcke 2013). Ob ihres stark formalisierten Charakters operieren netzwerkanalytische Konzeptualisierungen dabei zunächst mit epistemischen Objekten, die sich erheblich von den Objekten der ›klassischen‹ Literaturwissenschaft unterscheiden. Gerade deshalb aber bilden solche Konzeptualisierungen ein ebenso attraktives wie kontroverses Experimentierfeld für computerbasierte Zugänge zum Gegenstandsbereich ›Literatur‹, die nicht nur neue Antworten auf alte Fragen finden, sondern dezidiert *andere* Fragen formulieren wollen. Diese Ausrichtung wird noch unterstützt durch die *distant reading*-Affinitität der literaturwissenschaftlichen Netzwerkanalyse: Zwar lässt sich die visuelle Auswertung in Form von statischen oder dynamischen Netzwerkgraphen noch im Sinne des ›traditionellen‹ Paradigmas der Einzeltextanalyse verwenden (vgl. Moretti 2011); insbesondere die Auswertung von Netzwerkdaten mittels statistischer Methoden zielt jedoch auf die vergleichende Analyse größerer Korpora, die im Bereich der digitalen Dramenanalyse etwa mit historiographischem (Fischer et al. 2015) oder typologisierendem (Trilcke et al. 2016) Erkenntnisinteresse betrieben wird. Der hohe Abstraktionsgrad insbesondere der statistischen Ergebnisse von literaturwissenschaftlichen Netzwerkanalysen sowie deren Korpusorientierung führen allerdings zu einer Spannung zu ›traditionellen‹ Analyse- und Interpretationspraktiken der Literaturwissenschaft, mit denen die Ergebnisse der Netzwerkanalyse auf den ersten Blick schwer zu vermitteln sind. Hier zeigen sich gleichermaßen die Gefahren (ein Transfer der Ergebnisse zwischen digitalen Methoden und ›traditioneller‹ Literaturwissenschaft wird unmöglich) wie die Potenziale (die ›andersartigen‹ Ergebnisse der digitalen Methoden führen zur produktiven Irritationen der ›traditionellen‹ Literaturwissenschaft) der Methode, die in diesem Einzelvortrag anhand der Netzwerkanalyse von Dramen aus dem dlina-Korpus [1] exemplarisch diskutiert werden sollen.

Analyse der Figurenrede (Nils Reiter, Marcus Willand)

Computerlinguistische Methoden wie Named Entity Recognition und Koreferenzresolution (cf. Poesio et al. 2016) erlauben die Erkennung von Figurenreferenzen in der Rede dramatischer Figuren. Die erkannten Referenzen wiederum können genutzt werden, um den Stellenwert einer Figur innerhalb des Gesamttextes zu identifizieren. Neben der direkten Präsenz von Figuren (im Sinne von: Figur spricht; siehe auch das Problem der sog. Konfiguration, hierzu Ilsemann 1995, 2008) lässt sich damit auch die indirekte Präsenz (über eine Figur wird gesprochen) messen.

Im Falle von *Miss Sara Sampson* und *Emilia Galotti* (Lessing 1755, 1772) unterscheiden sich die beiden titelgebenden Figuren – Sara und Emilia – hinsichtlich dieser Dimensionen: Während Sara den größten Redeanteil auf sich vereinigt, spricht Emilia weniger als halb so viel (relativiert für die Länge des Gesamttextes) [2] . Im Gegensatz dazu wird *über* Emilia viel öfter gesprochen, so dass sie sozusagen passive Bühnenpräsenz zeigt. Anhand von Figuren wie dem *König* zeigt sich, dass auch passive Figuren die dramatische Handlung beeinflussen können. Dies gilt auch für Figuren und figurenähnliche Entitäten, die nicht in den Dramatis Personae genannt werden (*Gott*, das *Volk*).

Unser Beitrag zum Panel diskutiert zum einen die Herausforderungen an die maschinelle Sprachverarbeitung, wenn sie auf Dramentexte angewendet wird (Blessing et al. 2016). Zum anderen wollen wir untersuchen, inwiefern Autorinnen und Autoren sprachliche Eigenheiten der Figuren nutzen, um diese zu charakterisieren und z.B. als bestimmten Figurentypus (zärtlicher Vater, Hanswurst usw.; cf. Sørensen 1984, Aust 1989, Kord 2009) zu kennzeichnen.

Bilanzierung, Konsolidierung, Agenda

Die unterschiedlichen methodischen Zugänge zu dramatischen Texten erlauben zwar eine direkte Gegenüberstellung und Diskussion der drei Forschungsansätze, ihrer Prämissen, aber auch der Relevanz ihrer Ergebnisse für literaturtheoretische oder -historische Fragestellungen. Die vorgestellten Verfahren sollen letztlich aber nicht als konkurrierend oder unverbunden gedacht werden, sondern als Beiträge zu einem gemeinsamen Ziel: dem differenzierteren literaturwissenschaftlichen Verständnis dramatischer Texte. Vor dem Hintergrund der das Panel leitenden Idee einer Bilanzierung bisheriger und Konsolidierung aktueller Forschung auf dem Gebiet der *Digitalen Dramenanalyse* könnten ausgehend von den Einzelbeiträgen daher folgende Fragen diskutiert werden:

- Jede der drei Methoden verfolgt spezifische Fragen und birgt spezifische Herausforderungen. In welchem Maße gibt es gemeinsame Forschungsziele, zu denen jede der Methoden einen Beitrag leisten kann? Können die verschiedenen Methoden beispielsweise einen Beitrag zu einer empirisch gesicherten Gattungsdifferenzierung oder für die literaturgeschichtliche Periodisierung leisten?
- Wie können Ergebnisse, die mit unterschiedlichen methodischem Vorgehen gewonnen wurden, in Bezug zueinander gesetzt werden?
- Welche Ressourcen (insbesondere Textsammlungen) liegen vor und wie kann die Verfügbarkeit geeigneter Ressourcen für die Digitale Dramenanalyse zukünftig verbessert werden? Wie können die teils unterschiedliche Anforderungen der Methoden an die Formate von Daten und Metadaten aufgefangen werden?
- Welche konzeptuellen und datenbezogenen Standards für dokumentbezogene Metadaten und strukturelle oder semantische, lokale Annotationen liegen vor, wie kann die Standardisierung (bspw. durch Annotationsrichtlinien) weiter gefördert werden?
- Welche Tools sind für die digitale Dramenanalyse derzeit verfügbar, wie könnte die Tool-Entwicklung zielgerichtet gefördert werden? Welche generischen Tools könnten produktiv eingesetzt werden, wie könnte der Einsatzbereich vorhandener Tools erweitert (Adaptierbarkeit, Übertragbarkeit) und so eine breitere Nutzerbasis geschaffen werden?

Indem das Panel die Vielfalt digitaler Dramenanalysen vorführt und die explorative Kraft methodischer Innovation durch die Digital Humanities für die Literaturwissenschaften betont, möchten wir die fingierte "Laborsituation" im Sinne der theoretischen und wissenschaftspolitischen Implikationen einer auf Überprüfbarkeit und Wiederholbarkeit angelegten Wissenschaft verstanden wissen.

Fußnoten

1. https://dlina.github.io/Introducing-DLINA-Corpus-15-07-Codename-Sydney/
2. https://quadrama.github.io/blog/2016/10/07/ottokar-capulet

Bibliographie

Aust, Hugo (1989): *Volksstück vom Hanswurstspiel zum sozialen Drama der Gegenwart*. München: Beck.

Blei, David M. (2012): „Probabilistic Topic Models“, in: *Communication of the ACM* 55 (4): 77–84 10.1145/2133806.2133826.

Blessing, Andre / Bockwinkel, Peggy / Reiter, Nils / Willand, Marcus (2016): „Dramenwerkbank: Automatische Sprachverarbeitung zur Analyse von Figurenrede“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 281–284 http://dhd2016.de/boa.pdf [letzter Zugriff 24. August 2016].

Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario / Trilcke, Peer (2015): „Digital Network Analysis of Dramatic Texts“, in: *DH2015: Global Digital Humanities* http://dh2015.org/abstracts/xml/FISCHER_Frank_Digital_Network_Analysis_of_Dramati/FISCHER_Frank_Digital_Network_Analysis_of_Dramatic_Text.html [letzter Zugriff 24. August 2016].

Ilsemann, Hartmut (1995): „Computerized Drama Analysis“, in: *Literary and Linguistic Computing* 10 (1): 11–21.

Ilsemann, Hartmut (2008): „More statistical observations on speech lengths in Shakespeare’s plays“, in: *Literary and Linguistic Computing* 23 (4): 397–407.

Kord, Susanne (2009): „Unmöglichkeiten. Vater-Tochter-Dramen im 18. und 19. Jahrhundert“, in: Martinec, Thomas / Nitschke, Claudia (eds.): *Familie und Identität in der deutschen Literatur*. Frankfurt am Main: Peter Lang 105–126.

Moretti, Franco (2011): „Network Theory, Plot Analysis“, in: *Stanford Literary Lab Pamphlets* 2 http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf [letzter Zugriff 24. August 2016].

Poesio, Massimo / Stuckardt, Roland / Versley, Yannick (2016): *Anaphora Resolution: Algorithms, Resources, and Applications*. Berlin / Heidelberg: Springer.

Schöch, Christof (2016): „What Are Literary Topics, Really?“, in: *Digital Humanities Lunch*. Krakau, Institut für Polnische Sprache, 8. April 2016 http://christofs.github.io/literary-topics/#/ [letzter Zugriff 24. August 2016].

Schöch, Christof (im Erscheinen): „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama“, in: *Digital Humanities Quarterly*. (Preprint): https://zenodo.org/record/166356 [letzter Zugriff 15. November 2016].

Sørensen, Bengt Algot (1984): *Herrschaft und Zärtlichkeit der Patriarchalismus und das Drama im 18. Jahrhundert*. München: C.H. Beck.

Titzmann, Michael (1977): *Strukturale Textanalyse: Theorie und Praxis der Interpretation*. München: W. Fink.

Trilcke, Peer / Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario (2016): „Theatre Plays as ‚Small Worlds‘? Network Data on the History and Typology of German Drama, 1730–1930“, in: *Digital Humanities 2016: Conference Abstracts* http://dh2016.adho.org/abstracts/360 [letzter Zugriff 24. August 2016].

Trilcke, Peer (2013): „Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft“, in: Ajouri, Philip / Mellmann, Katja / Rauen, Christoph (eds.): *Empirie in der Literaturwissenschaft*. Münster: mentis 201–247.

Wasserman, Stanley / Faust, Katherine (1994): *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

Citizen Science unter dem Blickwinkel nachhaltiger sozialer und technischer Infrastrukturen

Seltmann, Melanie
melanie.seltmann@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Wandl-Vogt, Eveline
eveline.wandl-vogt@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Dorn, Amelie
amelie.dorn@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Unter Citizen Science wird eine Forschungsform verstanden, in der Wissenschaftler.Innen von Citizens in verschiedenen Bereichen des Forschungsprozesses unterstützt werden oder die gänzlich von Citizens durchgeführt wird. Die Verwendung von Citizen Science bringt für beide Seiten gewisse Vorteile: Wissenschaftler.Innen können auf neue Aspekte in ihrem Forschungsschwerpunkt kommen, und Citizens können den Elfenbeinturm der Wissenschaft durchbrechen und sich mit ihren eigenen Interessen in den Forschungsprozess einbringen (vgl. Riesch & Potter, 2014). Zudem kommt es häufig zur Community-Bildung innerhalb von Interessensgruppen, so darf der soziale Aspekt nicht vernachlässigt werden.

Während Citizen Science ursprünglich oft mit Arbeitsweisen in naturwissenschaftlichen Bereichen in Verbindung gebracht wurde, ist die Rolle von Citizen Science in den Humanities im Vergleich dazu noch weniger ausführlich beleuchtet und scheint erst in den letzten Jahren zunehmend an Aufmerksamkeit, zum Beispiel durch gezielte Förderprogramme oder Infrastrukturen, zu gewinnen. Aber ist dem tatsächlich so? Dabei bietet die Hinzunahme der Citizen Science in Forschungsprojekte ein großes Potential in verschiedenen Bereichen: begonnen vom Finden interessanter und gesellschaftsrelevanter Fragestellungen über die den meisten wohl zuerst in den Sinn kommende Datensammlung bis hin zur Disseminierung können Citizens zentrale Rollen, wie beispielsweise Partizipation bei der Entwicklung von Forschungsfragen, Co-Design des Forschungsprozesses oder der Resultate, einnehmen. Welche Rolle spielt Citizen Science demnach nun in den Humanities?

In dem vorgeschlagenen Panel greifen wir diese Thematik unter verschiedenen Ansatzpunkten auf. Es soll diskutiert werden, welche Rahmenbedingungen es für gute Citizen Science in (neuen) Humanities-Forschungsprojekten gibt und braucht. Zudem soll angeregt werden, eine Gemeinschaft für die Vernetzung und explorative Forschung im Rahmen der Digital Humanities aufzubauen und zu etablieren. In einem ersten Teil stellen die Panelisten ihre Standpunkte zum Thema Citizen Science in den Humanities dar. Diese vertreten unterschiedliche Plattformen, Fördergeber und führende Akteure in den Citizen Science. Namentlich handelt es sich um:

Celine Loibl *(Bundesministerium für Wissenschaft, Forschung und Wirtschaft (Österreich)): Programme Direktor Sparkling Science*

Das Förderprogramm des österreichischen BMWFW “Sparkling Science” fördert seit 2007 Citizen Science Projekte. Es setzt bereits im Schulalter an und bringt – einzigartig in Europa – Wissenschaftler.Innen und Schüler.Innen zusammen. Mit dieser Form von Citizen Science zielt es im Besonderen darauf ab, bei den Schüler.Innen Interesse an der Forschung zu wecken und somit Forschung und Bildungspolitik zu verbinden. Des Weiteren können durch die Zusammenarbeit von Wissenschaft und Schulen innovative wissenschaftliche Ansätze und Erkenntnisse generiert werden. Es soll erläutert werden, warum derartige Förderprogramme zur Verfügung gestellt werden. Weiters soll deren Situation in einem deutschsprachigen Kontext erschlossen werden.

Mike Mertens *(DARIAH): CEO, DARIAH Public Humanities Grant*

Nachdem auf den Istzustand der Verwendung von Citizen Science in den Humanities eingegangen wurde, soll das Augenmerk darauf gelegt werden, wie Infrastrukturen wie DARIAH Citizen Science sehen, und warum solche Organisationen die Hinzunahme von Citizen Science in den Humanities fördern. Warum hat Citizen Science in Förderprogrammen für die Humanities eine europäische Perspektive, und vor allem welche? Es soll erörtert werden warum dies wichtig ist und inwiefern sich dadurch etwas an unserer Forschung ändern könnte.

Fermin Serrano Sanz *(Socientize/Institute for Biocomputation and Physics of Complex Systems (BIFI), Univ. Zaragoza; Responsible Research and Innovation (RRI) liaison at ECSA): Citizen Science Infrastrukturen*

Ein wichtiger Punkt, damit Citizen Science so eingesetzt werden kann, wie von allen Beteiligten (Fördergeber, Wissenschaftler und Citizens) gewünscht, ist die Verwendung zuverlässiger und leicht verständlicher und nutzbarer Infrastrukturen. Es soll einen Überblick über Möglichkeiten von etablierten Infrastrukturen geben, sowie deren Nutzen insbesondere für die Humanities aufgezeigt werden. Es wird erläutert inwiefern Infrastrukturen. u.A. auch aus technischer Sicht, zur Verfügung stehen und inwiefern diese von Citizen Science Projekten genutzt werden, beziehungsweise nutzbar sind. Als Beispiele

können das Projekt Socientize, das Weißbuch Citizen Science, aber ebenso Themen wie "do-it-yourself" und "responsible science", sowie die soziale Infrastruktur ECSA und die COST Action on Citizen Science herangezogen werden.

Roberto Barbera *(Univ. of Catania/National Institute for Nuclear Physics (INFN)): From Open Access and Open Data to Open Science*

Schließlich soll der Blick geöffnet werden. Wie funktioniert Citizen Science in anderen Disziplinen und wie kann eine gute Open Access und Open Data Policy zu nachhaltiger und guter Open Science führen. Es soll erörtert werden inwiefern Open Science Commons (ein Ansatz zur gemeinsamen Nutzung digitaler Dienste, wissenschaftlicher Instrumente, Daten, Wissen, etc. für leichtere und produktivere Zusammenarbeit) Citizen Science einbringen kann und inwiefern aktuelle Forschungsinfrastrukturen Citizen Science unterstützen.

Eveline Wandl-Vogt *(Österreichische Akademie der Wissenschaften): Koordinatorin des Lexicography Laboratory (lexlab) ACDH, Koordinatorin des Projekts exploreAT!*

Sie berichtet von ihren Erfahrungen im Bereich Open Innovation, Open Science und Citizen Science, sowie Open Innovation in Science Methoden und deren Bedeutung am Beispiel eines konkreten Projektes, *exploreAT!*. Weiters wird der Citizen Science Survey, der im Rahmen des Projektes *exploreAT!* entstanden ist, vorgestellt, um einen Status quo der Citizen Science in Humanities Projekten abschätzen und darauf basierend Empfehlungen für (zukünftige) Projekte geben zu können.

Die Moderation des Panels wird durchgeführt von **Amelie Dorn** (*Österreichische Akademie der Wissenschaften*; Projektmitarbeiterin *exploreAT!*).

Im zweiten Teil werden im Podium wesentliche Fragen zu den Citizen Science in den (Digital) Humanities diskutiert. Begonnen wird mit der Metaebene, inwiefern Wissen etwas vergleichbares wie das *Mode 2* im Sinne einer *Knowledge Society* (vgl. Gibbons et al., 1994 und Nowotny et al., 2003) ist, nämlich nicht ein öffentliches Gut, sondern ein geistiges Eigentum, das wie andere Güter und Dienstleistungen in einer Knowledge Society produziert, angehäuft und gehandelt wird. Ebenso soll gefragt werden, wie sich die Wissenschaft in ein solches Modell einbetten kann.

Im zweiten Schritt soll erörtert werden, wie eine Förderung der Citizen Science aussieht oder aussehen kann. Daraufhin sollen politische und technische Infrastrukturen unter denen Citizen Science stattfinden kann diskutiert werden. Von den politischen und technischen Infrastrukturen soll auf die sozialen Infrastrukturen übergegangen werden. Es soll beleuchtet werden, warum Fördertöpfe von Nöten sind und ob diese eher für Citizens oder Wissenschaftler.Innen gebraucht werden. Zudem soll die Frage aufgeworfen werden, ob es damit weiters nur zu wissenschaftsgetriebener Citizen Science kommen kann. Es stellt sich die Frage, ob mit Fördertöpfen wirklich der Aufbau von Netzwerken gefördert wird.

In einem weiteren Block werden Ein- und Ausblicke von Fördergebern wie Sparkling Science und Infrastrukturen wie DARIAH erörtert, warum Citizen Science *in* ist, warum Citizen Science verwendet werden sollte und warum gerade zum jetzigen Zeitpunkt. Ebenso soll auf Strategiepapiere wie das White Paper (Serrano et al., 2016) auf europäischer Ebene sowie das Grünbuch (Bonn et al., 2016) auf deutscher Ebene und deren konkreten Nutzen eingegangen werden. Im Hinblick auf diese Papiere soll gefragt werden, was zum einen die Politik als ein gutes Projekt bezeichnet, zum anderen was der Wissenschaftler und inwiefern sich die beiden Ansichten voneinander unterscheiden. Am Beispiel von Österreich soll verdeutlicht werden, inwiefern der Support für Wissenschaftler zu mehr Forschungsprojekten geführt hat.

Schließlich sollen Infrastrukturen und Interakteure miteinander verbunden werden. Was ist die konkrete Rolle der technischen Infrastrukturen? Können diese den Prozess der Projekte verbessern, beschleunigen und/oder festigen? Wie können sie zu einer gediegenen Projektentwicklung beitragen? Und werden die technischen Infrastrukturen nur von den Wissenschaftler.Innen genutzt oder auch von den Laien? Beziehungsweise werden sie überhaupt genutzt?

Nachdem die Panelisten über diese Fragestellungen diskutiert haben, wird im dritten Teil das Publikum eingeschlossen. Zum einen gibt es vorbereitete Befragungen an das Publikum zu den zuvor diskutierten Inhalten, zum anderen kann das Publikum eigene aufgekommene Fragen an das Podium stellen. Hierbei soll unter anderem auf die Einschätzungen des Publikum zur Verwendung von Citizen Science konkret in den Humanities eingegangen werden. Die Annahmen des Publikums sollen mit den Ergebnissen aus dem Citizen Science Survey verglichen werden und Fragen des Publikums damit zu beantworten versucht werden.

Des Weiteren soll erörtert werden, wie ein positives Bild von technischer und sozialer

Entwicklung aussehen kann, durch welches für alle Seiten ermöglicht werden kann, bessere Projekte hervorzubringen. Es wird versucht, diese prinzipielle Frage durch eine interaktive Direktbefragung zu entschlüsseln. Daraufhin wird das Podium gebeten, hierauf eine Antwort zu finden. Anschließend soll die Frage für das Publikum geöffnet werden.

Abschließend lässt sich zusammenfassen, dass das vorgeschlagene Panel wichtige Themen sowie einen vielfältigen, breiten Einblick in die sozialen und technischen Strukturen der Citizen Science anspricht und diskutiert. Die Ergebnisse der Diskussion zwischen Panel und Podium, zwischen Citizens und Wissenschaftlern, bieten die Möglichkeit für weitere Ansatzpunkte um einen gemeinsamen Weg für zukünftige Ausrichtungen zu schaffen. Ausserdem bietet das vorgeschlagene Panel die Möglichkeit Personen aus den verschiedensten Bereichen die mit Citizen Science in Berührung kommen im Dialog zusammenzubringen, was nicht nur eine günstige Chance für potentielle Weiterentwicklungen im Citizen Science Bereich in den Humanities ist, sondern auch unabdingbar ist, um den direkten Austausch und Kontakt auf lokaler und weiterer europäischer Ebene zu stärken sowie zu avancieren.

Bibliographie

Bonn, Aletta / Richter, Anne / Vohland, Katrin / Pettibone, Lisa / Brandt, Miriam / Feldmann, Reinart / Goebel, Claudia / Grefe, Christiane / Hecker, Susanne / Hennen, Leonhard / Hofer, Heribert / Kiefer, Sarah / Klotz, Stefan / Kluttig, Thekla / Krause, Jens / Küsel, Kirsten / Liedtke, Christin / Mahla, A. / Neumeier, V. / Premke-Kraus, Matthias / Rillig, M. C. / Röller, Oliver / Schäffler, Livia / Schmalzbauer, Bettina / Schneidewind, Uwe / Schumann, Anke / Settele, Josef / Tochtermann, Klaus / Tockner, Klement / Vogel, Johannes / Volkmann, Wiebke / von Unger, Hella / Walter, D. / Weisskopf, Markus / Wirth, Christian / Witt, Thorsten / Wolst, Doris / Ziegler, David (2016): *Grünbuch Citizen Science Strategie 2020 für Deutschland.* Leipzig: Helmholtz-Zentrum für Umweltforschung (UFZ) / Deutsches Zentrum für integrative Biodiversitätsforschung (iDiv) Halle-Jena-Leipzig / Berlin: Museum für Naturkunde Berlin / Leibniz-Institut für Evolutions- und Biodiversitätsforschung (MfN) / Berlin-Brandenburgisches Institut für Biodiversitätsforschung (BBIB). http://www.buergerschaffenwissen.de/sites/default/files/assets/dokumente/gewiss-gruenbuch_citizen_science_strategie.pdf [letzter Zugriff 01. Dezember 2016].

Gibbons, Michael / Limoges, Camille / Nowotny, Helga / Schwartzman, Simon / Scott, Peter / Trow, Martin (1994): *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. London: Sage.

Nowotny, Helga / Scott, Peter / Gibbons, Michael (2003): „Mode 2 revisited: The New Production of Knowledge“, in: *Minerva* 41: 179–194.

Riesch, Hauke / Potter, Clive (2014): „Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions“, in: *Public Understanding of Science* 23 (1): 107–120.

Serrano Sanz, Fermín / Holocher-Ertl, Teresa / Kieslinger, Barbara / Sanz García, Francisco / Silva, Cândida G. (2014): *White Paper on Citizen Science for Europe*. Socientize Consortium https://www.zsi.at/object/project/2340/attach/White_Paper-Final-Print.pdf [letzter Zugriff 01. Dezember 2016].

Das digitale Museum: ein nachhaltiger Partner der Digital Humanities?

Hohmann, Georg
g.hohmann@deutsches-museum.de
Deutsches Museum, München

Schmidt, Antje
antje.schmidt@mkg-hamburg.de
Museum für Kunst und Gewerbe, Hamburg

Doppelbauer, Regina
r.doppelbauer@albertina.at
Albertina, Wien

Rehbein, Malte
malte.rehbein@uni-passau.de
Universität Passau

Ausgangslage

In der digitalen Gesellschaft steht das Museum als Gedächtnisinstitution vor besonderen Herausforderungen. Sammeln, Bewahren, Forschen und Vermitteln als die

klassischen Aufgabenbereiche des Museums müssen von Grund auf hinterfragt und in Hinblick auf digitale Möglichkeiten und Anforderungen angepasst, verändert und erweitert werden.

Wie kaum ein anderes Fach steht die Disziplin der Digital Humanities als Repräsentant dafür, welche Anforderungen und Ansprüche vor allem von Seiten der Wissenschaft an moderne Gedächtnisinstitutionen gestellt werden. Waren und sind Museen seit jeher wichtige Partner und auch Horte der historisch orientierten Geisteswissenschaften, müssen sie sich nun an die Gegebenheiten der digitalen Geisteswissenschaften anpassen, um ihre Bedeutung zu erhalten und vielleicht sogar auszubauen. Das Museum kann nicht mehr nur ein begehbarer Ort des kulturellen Erbes sein, sondern muss auch als digitaler Wissensspeicher seine Informationen zur Verfügung stellen (Clough 2013: 2). Die Verantwortung gegenüber den realen Objekten erweitert sich auf die Sphäre der digitalen Information (Keene 1998: 23), die ebenso wie diese nach wissenschaftlichen Maßstäben gesammelt, bewahrt, erforscht und vermittelt werden will, wobei das Museum gleichzeitig als Aggregationspunkt und Erzeuger fungiert. Dabei wird das Museum an seinen eigenen Ansprüchen gemessen, nämlich die Objekte seiner Betrachtung „für die Ewigkeit“ zu bewahren und dauerhaft der Allgemeinheit zur Verfügung zu stellen (ICOM 2004). Wie kann dies in einer digitalen Umgebung nachhaltig gelingen?

Themenbereiche

Digitalisierung

Die Digitalisierung von Objekten hat inzwischen flächendeckenden Einzug in die Museen gehalten. Grundsätzlich geht es dabei um den Vorgang, von analogen Objekten digitale Abbilder zu generieren, aber auch um die Überführung analoger Informationsträgern in digitale Formate. Hier sind eine Vielzahl von grundlegenden Entscheidungen zu treffen. Dazu gehört die Definition des eigenen Qualitätsanspruchs und dessen Abwägung mit den Anforderungen einer ökonomischen Massendigitalisierung. Durch die ständige Weiterentwicklung im Bereich der digitalen Erfassungstechniken stellt sich auch die Frage, ob eine schlichte Digitalfotografie zur Digitalisierung überhaupt ausreichend ist. Soll ein Digitalisat ein Original in der Ausstellung und für Forschungsfragen gar ersetzen, um das Original besser zu schützen?

Erschließung

Ein Aspekt der Digitalisierung, der oft subsumiert wird, ist die wissenschaftliche Erschließung von Objekten zur Erzeugung digitaler Daten. Im Museum ist dies oft kein einfacher Vorgang der Übertragung von analoger in digitale Information, da die Information im Vorfeld gar nicht strukturiert vorhanden ist, sondern erst wissenschaftlich erarbeitet werden muss. Die Tiefe und die Perspektive mit der digitalisiert und erschlossen wird, bestimmen zu einem nicht geringen Grad die Möglichkeiten der späteren wissenschaftlichen Bearbeitung. In der Praxis ist die Erschließung häufig eher von pragmatischen Erwägungen geprägt als von konzeptuellem Vorgehen (Koch 2015). Auch stellt sich die Frage nach dem Umfang. Werden einzelne Objekte in der Tiefe erschlossen, oder wird – in einem ersten Schritt - auf eine flächendeckende Flacherschließung gesetzt, um die Quantität (mit Verlust der Qualität) zu steigern? Nicht zuletzt sind die Kenntnis und die Anwendung von adäquaten Standards, Normdaten und Techniken für eine nachhaltige Erschließung unabdingbar.

Langzeitarchivierung und -verfügbarkeit

Die größte technologische Herausforderung stellt sich im Bereich der digitalen Langzeitarchivierung. Mit ihren digitalen Assets sollten Museen ebenso wie mit ihren realen Objekten umgehen, d.h. sie dauerhaft archivieren und die idealen Lagerbedingungen einhalten. Aufgrund der schieren Menge an digitalen Daten kann diese Aufgaben kaum mehr adäquat von einzelnen Häusern alleine gestemmt werden. Auch die notwendige technische Expertise zur Einhaltung internationaler Standards zur Langzeitarchivierung übersteigt Fähigkeiten und Aufgaben eines Museums. Hier bieten sich Kooperationen mit Dienstleistern an, die aber koordiniert und nachhaltig finanziert werden wollen.

Rechtliche Rahmenbedingungen

Spätestens bei der Veröffentlichung von Daten stellt das Rechtemanagement eine große Hürde dar (Müller, Truckenbrodt 2013). Welche Daten oder Digitalisate dürfen überhaupt gezeigt werden? Wie ist die Rechteklärung zu organisieren und welche verschiedenen Gesetze (Urheberrecht, Markenrecht, Vervielfältigungsrecht, Nutzungsrecht etc.) sind zu berücksichtigen? Die Prinzipien des Open Access sind auch für viele Museen erstrebenswert, aber die weitgehend unklare Rechtslage ist letztlich oft ein Hinderungsgrund, die Open Access Gedanken vollständig umzusetzen (Hamburger Note 2015).

Bereitstellung und Vermittlung

Die digitalen Inhalte von Museen werden üblicherweise über Online-Portale vermittelt. Um aber explizit eine wissenschaftliche Nutzung der Daten zu ermöglichen, sind zudem viele Rahmenbedingungen einzuhalten und Schnittstellen anzubieten. Besonders die digitalen Geisteswissenschaften stellen in dieser Hinsicht hohe Anforderungen, um automatisierte Verfahren anwenden zu können. Die Daten sollten im Idealfall über technische Schnittstellen verfügbar sein, die wieder selbst nach unterschiedlichen Vorgaben und Modellen organisiert sein können. Zudem kann in diesem Bereich noch kaum auf etablierte Systeme zurückgegriffen werden, wodurch Eigenentwicklungen notwendig werden.

Strukturwandel

Mit einer Hinwendung zum Digitalen geht auch eine Metamorphose der Institution Museum einher, die Auswirkungen auf Arbeitsabläufe, Aufgabenbereiche und Zielsetzungen hat. Dies kann sogar so weit führen, das ganze Aufgabenbereiche wegfallen und und andernorts neue entstehen, etwa die eines Museum Information Curators (Low, Doerr 2010). Um den entsprechenden Nachwuchs heranzubilden, sind Museen zur aktiven Mitgestaltung der Aus- und Weiterbildung aufgefordert. Diese tiefgreifenden strukturellen Änderungen sollten von einem umfassenden Change Management begleitet werden, was aber die Kapazitäten der meisten Häuser übersteigen dürfte.

Perspektiven

Das Panel widmet sich den skizzierten Themenbereichen unter verschiedenen Perspektiven, wobei von der These ausgegangen wird, dass die Themen eng zusammenhängen und aufeinander aufbauen. Es kann daher weniger um eine Detaildiskussion einzelner Techniken und Vorgehensweisen gehen, sondern um die Feststellung des State-of-the-Art in der deutschsprachigen Museumsszene sowie um die Diskussion der Zukunft von Museen in Hinblick auf ihr Potenzial als nachhaltiger Partner für die digitale Gesellschaft und Forschung.

Impulsvorträge

Moderation: Mareike Schumacher (Universität Hamburg) & Etta Grotrian (Jüdisches Museum Berlin)

Georg Hohmann: Das digitale Museum

In einer umfassenden Maßnahme werden am Deutschen Museum die Bestände aus Objektsammlungen, Archiv und Bibliothek erschlossen und digitalisiert, womit der Weg zu einem Digitalen Museum eingeleitet wird. Die Ergebnisse werden in einem gemeinsamen Online-Portal präsentiert, das den Wissenskosmos des Deutschen Museum sowohl für wissenschaftliche als auch für interessierte Fachnutzer in aller Welt zugänglich macht. Ein großes Potential hat die interne und externe Vernetzung der Daten, bei der die Nutzung einheitlicher Standards und Normdaten eine zentrale Rolle spielt. Der Beitrag fokussiert die technischen Aspekte zur Bereitstellung von musealen Forschungsdaten und thematisiert die Voraussetzungen und Perspektive zur Nutzung dieser Daten in den digitalen Geisteswissenschaften.

Antje Schmidt: Offene Daten als nachhaltige Ressource

Die Herausforderungen liegen für die Museen heutzutage nicht nur in der digitalen

Bereitstellung von Informationen z.B. über Sammlungsdatenbanken, sondern auch in der nachhaltigen Vermittlung und Nachnutzbarmachung dieser. Das Management der rechtlichen Bedingungen, unter denen diese Informationen bereitgestellt werden können und deren klare Vermittlung sind dafür unabdingbar. Mit der MKG Sammlung Online hat das Museum für Kunst und Gewerbe Hamburg als erstes Museum in Deutschland diejenigen digitalisierten Bestände, für die dies rechtlich möglich ist, zur freien Nutzung zur Verfügung gestellt und dies mit Hilfe von Creative Commons Lizenzen dargestellt. Jedes einzelne Digitalisat kann individuell lizensiert werden. In den meisten Online-Sammlungspräsentationen sind die rechtlichen Metadaten allerdings an den Datensatz gebunden. Dies führt zu Problemen, wenn z.B. für ein Objekt mehrere Abbildungen mit unterschiedlichen Lizensierungen vorhanden sind. Zudem sind diese rechtlichen Metadaten nicht einheitlich mitgeführt, sobald es um die Weitergabe an andere Portale geht.

In dem Vortrag soll erläutert werden, welche Bedingungen geschaffen werden müssen, um das Potenzial digitaler Sammlungen zu entfalten, diese nachhaltig zu öffnen und nachnutzbar zu machen.

Regina Doppelbauer: Digitalisierung von 1400 Klebebänden

Der überwiegende Teil der Druckgraphiken der Albertina Wien ist in historischen Großfoliobänden eingeklebt. Diese 1436 Volumina spiegeln Wissen und Ästhetik des 18. und frühen 19. Jahrhunderts wider. Die Blätter selbst erzählen die Entwicklung der Druckgraphik und enthalten ikonographisch unser neuzeitliches Bildgedächtnis. Ein Forschungsprojekt der Albertina zielt auf die dringend notwendige Autopsie und Veröffentlichung der Bände: Diese werden digital erfasst, mit Metadaten versehen und so rasch wie möglich nicht nur der Forschungscommunity online zur Verfügung gestellt.

Der Beitrag stellt den Ansatz vor, der Fülle mit Augenmaß zu begegnen und gleichwohl Standards und Nachhaltigkeit zu gewährleisten: Da eine Bandseite bis zu zwanzig Objekte aufweist, ist eine Einzelobjekterfassung von geschätzten 500.000 Werken nicht zu leisten. Es wird daher ein generisches Erfassungsmodell entwickelt, das vom obligatorischen Scan jeder Seite bis hin zu einer detaillierten Metadatenerfassung der darauf montierten Objekte mehrere Stufen der Erschließung ermöglicht. Wird ein Band flach erschlossen, so werden alle technischen Vorkehrungen getroffen, um spätere Anreicherungen – hausintern oder durch crowd/niche-sourcing - vornehmen zu können.

Malte Rehbein: Virtuelle Verbundsysteme als Nachhaltigkeitsstrategie für Museen und andere Kulturerbe-Institutionen

Sowohl für die Bewahrung des Kulturellen Erbes als auch für dessen Präsentation bietet die Digitalisierung neue Möglichkeiten; dass in der Regel erhebliche Ressourcen aufzuwenden sind, um diese Chancen des Digitalisierungstrends zu nutzen, ist vor allem für kleine und mittlere Institutionen eine große Herausforderung. Zudem ist eine nachhaltige Ausgestaltung der digitalen Innovationen ein Schlüssel für ihren langfristigen Nutzen.

Der Vortrag illustriert die Anforderungen an Museen aus der Sicht der Digital Humanities am Beispiel des 2016 gestarteten Projekts „Virtuelle Verbund-Systeme und Informations-Technologien für die touristische Erschließung von kulturellem Erbe (ViSIT)“, das in einem grenzüberschreitenden regionalen Verbund von Standorten und den dort ansässigen Kulturerbe-Institutionen mit Hilfe digitaler Kooperations- und Vermittlungsformen das Ziel verfolgt, die Vermittlung von Regionalgeschichte innovativ zu gestalten.

Bibliographie

Clough, G. Wayne (2013): *Best of both worlds: Museums, libraries, and archives in a digital age*. Washington: Smithsonian Institution.

Keene, Suzanne (1998): *Digital collections: Museums and the information age*. Oxford: Butterworth-Heinemann.

ICOM (2004): *ICOM Code of Ethics for Museums* http://icom.museum/the-vision/code-of-ethics/ [letzter Zugriff 24. August 2016].

Koch, Gertraud (2015): „Kultur digital. Tradieren und Produzieren unter neuen

Vorzeichen“, in: Bolenz, Eckard / Franken, Lina / Hänel, Dagmar: *Wenn das Erbe in die Wolke kommt: Digitalisierung und kulturelles Erbe*. Essen: Klartext.

Müller, Carl Christian / Truckenbrodt, Michael (2013): *Handbuch Urheberrecht im Museum: Praxiswissen für Museen, Ausstellungen, Sammlungen und Archive*. Bielefeld: transcript.

Hamburger Note (2015): *Hamburger Note zur Digitalisierung des kulturellen Erbes* http://hamburger-note.de/ [letzter Zugriff 24. August 2016].

Low, Jyue Tyan / Doerr, Martin (2010): *A Postcard is Not a Building: Why we Need Museum Information Curators* http://network.icom.museum/fileadmin/user_upload/minisites/cidoc/ConferencePapers/2010/low.pdf [letzter Zugriff 24. August 2016].

eValuation - Kriterien zur Evaluation digitaler Angebote und Forschungsinfrastrukturen

Kurmann, Eliane
eliane.kurmann@infoclio.ch
infoclio.ch, Schweiz

Baumann, Jan
jan.baumann@infoclio.ch
infoclio.ch, Schweiz

Natale, Enrico
enrico.natale@infoclio.ch
infoclio.ch, Schweiz

eingereicht von: infoclio.ch – Fachportal für die Geschichtswissenschaften der Schweiz

Referierende: Gabi Schneider / Alexander Hasgall / Philipp Steinkrüger

Moderation: Eliane Kurmann (infoclio.ch) / Jan Baumann (infoclio.ch)

Mit der Etablierung der Digital Humanities an den Universitäten und Forschungseinrichtungen werden, wenn auch noch zögerlich, neue Modelle zur Evaluation digitaler Infrastrukturprojekte und zur Rezension digitaler Inhalte und Angebote entwickelt und erprobt. Die Evaluationsverfahren dienen der Beurteilung der Projekte und Initiativen im Hinblick auf die weitere finanzielle Förderung und die Leistungsanerkennung im akademischen Umfeld. Beim Rezensieren geht es zudem um die Sichtbarmachung und Hervorhebung besonders gelungener Projekte und Angebote. Und schliesslich werden geplante Angebote und Infrastrukturen an den bereits etablierten Qualitätsmerkmalen ausgerichtet.

Verschiedene Institutionen und Vereinigungen sind damit beschäftigt, Evaluationsverfahren auszuarbeiten, die über die Messung der traditionellen Forschungsleistungen hinausgehen. Neben den fachspezifischen wissenschaftlichen Kriterien werden bei der Evaluation digitaler Angebote und Infrastrukturen beispielsweise auch technische Aspekte, die Interoperabilität, Design und Anwenderfreundlichkeit sowie das Interagieren von Inhalt und Präsentationsform, die Zugänglichkeit oder die Dauerhaftigkeit der Inhalte berücksichtigt.

Erste Kriterienkataloge sind bereits ausgestaltet, die Instrumente und Methoden ihrer Anwendung stellen weitere Herausforderungen dar: Wie verändern sich die Qualitätskriterien mit der fortlaufenden technischen Entwicklung? Wie wird etwa die Reichweite der Resultate im World Wide Web festgestellt und innerhalb der Forschungsevaluation gewertet? Was bedeutet Dauerhaftigkeit im digitalen Kontext? Diskutiert wird aber auch die grundsätzliche Frage, ob es angesichts der Vielfalt der digitalen Projekte überhaupt möglich ist, standardisierte Verfahren und einheitliche Richtlinien zu definieren. Und zielen die Neuerungen auf die Erweiterung der traditionellen Evaluationsverfahren, sodass diese auch auf digitale Projekte anwendbar werden, oder verlangen die Digital Humanities eigene Beurteilungsmodelle?

Für die Dhd2017-Tagung schlägt infoclio.ch ein Panel vor, in dem neue Evaluationsmodelle vorgestellt werden. Expertinnen und Experten, die sich mit der Konzipierung und Anwendung neuer Verfahren und Kriterien beschäftigen, berichten von ihren Erfahrungen und stellen grundsätzliche Überlegungen zur Diskussion. Die digitale Nachhaltigkeit wird dabei in zweifacherweise thematisiert: Zum einen wird sie als Qualitätsmerkmal in der Beurteilung von digitalen Inhalten, Tools und Infrastrukturen diskutiert. Zum andern fördert die Evaluation grundsätzlich die Qualität und damit die Nachhaltigkeit, da die professionelle Beurteilung eines Projekts seine Fortführung begünstigt.

Für die drei [1] nachfolgend beschriebenen Beiträge sind jeweils 10 bis 15 minütige Präsentationen vorgesehen, die zugleich

die Grundlage für die anschliessende Diskussion (45 Minuten) bilden. Die Beiträge beschäftigen sich mit der Evaluation von digitalen Forschungsinfrastrukturen, dem Umgang mit Digital-Humanities-Projekten in der Forschungsevaluation und mit der kritischen Besprechung von digitalen Editionen. Diskutiert werden bereits erprobte und im Entstehen begriffene Evaluationsverfahren, wobei die Erfahrungsberichte die Herausforderungen in der praktischen Anwendung deutlich machen. Mit Blick auf das Tagungsthema findet die digitale Nachhaltigkeit in allen Beiträgen besondere Beachtung. Einerseits soll thematisiert werden, welche Bedeutung der digitalen Nachhaltigkeit als Evaluationskriterium zukommt; zum andern sollen Erfahrungen aus der Praxis des Evaluierens zur Konkretisierung der Konzepte der *digitalen Nachhaltigkeit* beitragen. Gefragt wird unter anderem, wie die digitale Nachhaltigkeit „gemessen" wird, geht es dabei doch nicht nur um technische Aspekte, sondern auch um die freie Zugänglichkeit sowie die Nutzungsrechte der digitalen Inhalte und Infrastrukturen.

Beiträge

Gabi Schneider, stellvertretende Leiterin des Programms „ Wissenschaftliche Information: Zugang, Verarbeitung und Speicherung "

Beitrag: Evaluation digitaler Forschungsinfrastrukturen

Das Programm „Wissenschaftliche Information: Zugang, Verarbeitung und Speicherung" von swissuniversities fördert den Aufbau eines national verfügbaren Grundangebots an digitalen Inhalten sowie optimaler Werkzeuge (Tools und Infrastrukturen) für deren Verarbeitung. Projekte werden von den Hochschulen eingereicht und im Rahmen der Programmorganisation in Bezug auf die Mittelvergabe evaluiert. Die Qualität und die Nachhaltigkeit von Projekten werden in verschiedenen Stadien gefördert. Zum einen werden Kriterien wie technische Standards, Interoperabilität oder die Bezugnahme auf internationale Referenzprojekte in den Programmunterlagen (Strategiepapiere, Antragsformular und Wegleitung) explizit genannt. Zum anderen werden die Projekte im Evaluationsverfahren auf diese Kriterien hin geprüft. Im Rahmen des Programms wurde seit 2013 ein erstes Portfolio von Diensten aufgebaut. Im weiteren Verlauf sollen Anforderungskriterien für eine periodische Überprüfung dieser Dienste definiert werden. Da Grossprojekte mit „nationalem" oder internationalem Anspruch meistens von verschiedenen Geldgebern unterstützt werden, gewinnen dabei der Austausch und die Verständigung mit anderen Förderinstitutionen an Bedeutung. Der Beitrag zeigt Ansätze auf.

Alexander Hasgall, Wissenschaftlichen Koordinator, SUK P3 „ Perfomances de la recherche en sciences humaines et sociales "

Beitrag: Evaluationsverfahren in den Digital Humanities

Im Rahmen von Evaluationsverfahren spielen digital präsentierte Inhalte oftmals keine gesonderte Rolle. Jedoch weist die Forschung in den Digital Humanities u.a. im Hinblick auf Fragen von Zugänglichkeit, der Wahrnehmung und Verbreitung in der Wissenschaftscommunity oder auch der Nachhaltigkeit der Forschungsergebnisse wichtige Besonderheiten auf, welche in herkömmlichen Evaluationsverfahren nicht immer angemessen reflektiert werden. Im Rahmen des Panel-Beitrags soll auf die Auswirkung von Evaluationsverfahren auf die Forschung in den Digital Humanities eingegangen und zugleich diskutiert werden, inwieweit Nachhaltigkeit selbst ein Qualitätsmerkmal von Forschung bilden kann.

Philipp Steinkrüger, Gründungsmitglied des Instituts für Dokumentologie und Editorik und Managing Editor der Rezensionszeitschrift RIDE (Review Journal for digital editions and ressources)

Beitrag: Digitale Nachhaltigkeit im Kriterienkatalog

Die Zahl wissenschaftlicher Onlineangebote, darunter auch zahlreiche digitale Editionen, nimmt stetig zu. Eine kritische Reflektion und Evaluation solcher Angebote ist jedoch noch sehr peripher, da sich die etablierten Rezensionsorgane weiterhin auf Printpublikationen konzentrieren. RIDE, die erste Rezensionszeitschrift explizit für digitale Editionen, bietet seit 2015 ein Forum, in dem digitale Editionen kritisch besprochen werden. Der Komplexität solcher Editionen, die sich durch die vielfältigen Möglichkeiten des digitalen Paradigmas und ihrer Umsetzungen ergibt, versucht RIDE mit einem Kriterienkatalog zu begegnen, der Rezensenten in ihren Besprechungen leiten soll.

Der Beitrag wird insbesondere auf das Thema „digitale Nachhaltigkeit" eingehen. Erstens wird vorgestellt, was RIDE selbst zur digitalen Nachhaltigkeit beiträgt. Der Katalog als Grundlage aller Besprechungen enthält eine Reihe von Kriterien, die zentral für die Möglichkeit langfristiger Verfügbarkeit sind. Besprechungen in RIDE dokumentieren, ob

und inwiefern aktuelle digitale Editionen diese Kriterien erfüllen und tragen dazu bei, dass zukünftige Editionsprojekte diese Kriterien von Anfang an im Blick behalten. Zweitens wird für jede Besprechung eine Vielzahl von Aspekten formalisiert abgefragt und gespeichert. Dies erlaubt einen Einblick in die Frage, ob und wie aktuelle Editionen dem Thema „digitale Nachhaltigkeit“ begegnen. Obwohl das Sample noch zu klein ist, um eine allgemeingültige Aussage zu formulieren, gibt es doch Hinweise darauf, dass das Thema digitale Nachhaltigkeit noch mehr in den Fokus der Editorinnen und Editoren rücken muss, damit Editionen langfristig verfügbar gehalten werden können.

Richtlinien und Kriterienkataloge zur Evaluierung digitaler Projekte und Ressourcen

18thConnect; NINES: Guidelines and Peer Review Criteria. Online: 18thConnect – Eighteenth-century Scholarship < http://www.18thconnect.org/about/scholarship/peer-review/#new >

American Historical Association (2015): Guidelines for the Professional Evaluation of Digital Scholarship by Historians. Online: American Historical Association < https://www.historians.org/teaching-and-learning/digital-history-resources/evaluation-of-digital-scholarship-in-history/guidelines-for-the-professional-evaluation-of-digital-scholarship-by-historians >

Modern Language Association, Committee on Information Technology: Guidelines for Evaluating Work in Digital Humanities and Digital Media. Online: Modern Language Association < https://www.mla.org/About-Us/Governance/Committees/Committee-Listings/Professional-Issues/Committee-on-Information-Technology/Guidelines-for-Evaluating-Work-in-Digital-Humanities-and-Digital-Media >

Modern Language Association, Committee on Information Technology: Guidelines for Authors of Digital Resources. Online: Modern Language Association < https://www.mla.org/About-Us/Governance/Committees/Committee-Listings/Professional-Issues/Committee-on-Information-Technology/Guidelines-for-Authors-of-Digital-Resources >

Sahle, Patrick (2014): Kriterienkatalog für die Besprechung digitaler Editionen. Online: Institut für Dokumentologie und Editorik < http://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/ >

Fußnoten

1. Noch offen ist die Beteiligung des UsabilityLabs der HTW Chur, dem „Schweizer Kompetenzzentrum für die Evaluation von Online-Angeboten“. Der Beitrag würde sich auf die Erfahrungen des UsabilityLabs in der Evaluation digitaler Präsentationsformen richten.

Bibliographie

Akademien der Wissenschaften Schweiz (2014): *„Open Access“: Für einen freien Zugang zu Forschungsergebnissen*. Positionspapier der Schweizerischen Akademie der Medizinischen Wissenschaften. Swiss Academies Communications 9 (1) http://www.samw.ch/dam/jcr:9d2d13bd-1757-401a-962e-0a8ec946fb27/positionspapier_samw_open_access.pdf .

Arts and Humanities Research Council (2006): *Peer review and evaluation of digital resources for the arts and humanities*. Institute of Historical Research, School of Advanced Study, University of London http://www.history.ac.uk/sites/history.ac.uk/files/Peer_review_report2006.pdf .

European Strategy Forum on Research Infrastructures (2016): *Strategy Report on Research Infrastructures*. ESFRI Roadmap 2016 http://www.esfri.eu/sites/default/files/20160308_ROADMAP_single_page_LIGHT.pdf .

Open Scholar: *Independent Peer Review Manifesto* http://www.openscholar.org.uk/independent-peer-review-manifesto/ .

Pfannenschmidt, Sarah L. / Clement, Tanya E. (2014): „Evaluating Digital Scholarship: Suggestions and Strategies for the Text Encoding Initiative“, in: *Journal of the Text Encoding Initiative* 7 http://jtei.revues.org/949 .

DORA: *San Francisco Declaration on Research Assessment: Putting science into the assessment of research* http://www.ascb.org/dora/ .

Hackathons als Zukunftslabor für die digitale Nachhaltigkeit

Noyer, Frédéric
f.noyer@docuteam.ch
Verein Opendata.ch, Schweiz

Im vorgeschlagenen Panel gehen wir der Frage nach, welchen Mehrwert Hackathons für die nachhaltige Entwicklung der Digital Humanities bieten. Das Panel nimmt Bezug auf die Erfahrungen, die im Rahmen der beiden Open Cultural Hackathons gemacht wurden, die 2015 und 2016 von der schweizerischen OpenGLAM Working Group, einer Arbeitsgruppe des Vereins opendata.ch, und lokalen Partnern organisiert wurden.

Im Panel werden insbesondere drei Themen in Zusammenhang mit dem Tagungsthema Digitale Nachhaltigkeit angegangen. Das erste Thema betrifft die Rolle von Open Data im Forschungsbereich der Digital Humanities. Es soll aufgezeigt werden, in welchem Masse Open Data als nachhaltiger Faktor für die Digital Humanities-Forschung betrachtet werden kann, insbesondere in den Bereichen Qualitätssicherung und Wiederverwendung von digitalen Daten.

Das zweite Thema betrifft die Interdisziplinarität, die oft als essentieller Faktor einer gelungenen und nachhaltigen Ausbildung in den Digital Humanities betrachtet wird. Im Panel soll aufgezeigt werden, wie das Format des Hackathons eine Gelegenheit für die interdisziplinäre Auseinandersetzung zwischen verschiedenen Akteuren bieten kann, insbesondere zwischen Forschenden und Gedächtnisinstitutionen.

Beim dritten Thema, das ebenfalls im Zusammenhang mit der Entwicklung einer Ausbildung in den Digital Humanities steht, bezieht sich auf die praktische Ausrichtung des Hackathons. "Less Yak, more Hack" war eine der Devisen in den Anfängen der Digital Humanities. In welchem Masse kann ein experimenteller Rahmen, konzentriert auf eine kurze Zeitspanne, wie ihn die Hackathons zur Verfügung stellen, zu einer nachhaltigen Forschungsumgebung für die Digital Humanities beitragen?

Seit Jahren digitalisieren Bibliotheken, Archive und Museen ihre Bestände und machen diese teilweise online verfügbar. Parallel dazu liegen vermehrt auch Metadaten und andere strukturierte Daten in digitaler Form vor. Das Potential dieser Daten und Digitalisate ist heute jedoch bei weitem nicht ausgeschöpft. Hier setzt der Hackathon an, indem er Forschende, Vertreter von Museen, Archiven und Bibliotheken sowie Software-Entwicklerinnen, Wikipedianerinnen und Software-Designer zusammenbringt, damit diese in der Praxis gemeinsam das Potenzial des digitalen Kulturerbes und dessen Weiterverwendung ergründen können.

Ergebnisse des Hackathons bilden üblicherweise Projektideen, Konzepte und erste Prototypen sowie der Knowhow-Austausch und neue Kooperationen zwischen den Gedächtnisinstitutionen, den Forschenden und den anderen Teilnehmerinnen und Teilnehmern. Damit hat der Hackathon in erster Linie eine Katalysator-Funktion. Der Hackathon ist zudem eine Gelegenheit, Gedächtnisinstitutionen zu ermuntern, Kulturdaten zur freien Weiterverwendung bereitzustellen. Etliche Institutionen nutzen den Event, um einem interessierten Publikum neu verfügbare Daten zur weiteren Nutzung vorzustellen.

Das Panel besteht aus vier Beiträgen. Er wird die Prinzipien von Open Data vorstellen und aufzeigen, welche Rolle Hackathons aus Sicht der Organisierenden im Rahmen der Digital Humanities in der Schweiz spielen können. Die drei weiteren Beiträge wiederspiegeln die Sicht der drei verschiedenen Teilnehmergruppen: die Forschenden, die Gedächtnisinstitutionen sowie die Software-Entwickler und Webdesignerinnen. Gemeinsam werden sie über die Grundidee des Hackathons und deren konkrete Umsetzung diskutieren. Der Fokus liegt dabei auf Open Data und damit der langfristigen Bereitstellung von Daten, der Rolle von Open Source Software, der Austausch mit verschiedenen Stakeholdergruppen sowie der Vorgehensweise im Vorfeld und während des Anlasses. Konkret werden dabei die folgenden Fragen thematisiert:

- Welche Bedeutung hat die Open-Data-Politik für die digitale Nachhaltigkeit?
- Inwiefern ist das Format des Hackathons geeignet für die langfristige Entwicklung des Digital Humanities-Umfelds?
- Welche Rolle haben Hackathons bezüglich der längerfristigen Innovation und Etablierung von Qualitätsstandards?
- Die Hackathonprojekte zwischen Kurzlebigkeit und langfristigen Projekten und Lösungen: Wo liegt der Beitrag zur digitalen Nachhaltigkeit?

Frédéric Noyer ist Mitglied des Organisationsteams des Schweizer Kulturhackathons und wird im Rahmen des Panels die Perspektive von OpenGLAM und den Organisierenden des Hackathons vertreten.

OpenGLAM ist eine Bewegung der Open Knowledge Foundation, die sich an Gedächtnis- und Kulturinstitutionen richtet. Die Philosophie von Open(GLAM steht im Englischen für „Galleries, Libraries, Archives, Museums“) lässt

sich anhand von fünf Prinzipien einfach auf den Punkt bringen [1]:

1. Digitale Informationen zu Überlieferungsobjekten (Metadaten) werden mittels einer geeigneten Lizenz ohne Nutzungsbeschränkungen verfügbar gemacht […].

2. Gemeinfreie Werke werden (insbesondere im Zusammenhang mit der Digitalisierung) keinen neuen Nutzungsbeschränkungen unterworfen.

3. Bei der Publikation von Daten wird explizit und unmissverständlich kommuniziert, welche Art von Weiterverwendung erwünscht bzw. erlaubt ist […].

4. Bei der Publikation von Daten werden offene, maschinenlesbare Dateiformate verwendet.

5. Neue Möglichkeiten, Internet-NutzerInnen einzubeziehen, werden aktiv genutzt.

Damit wird für die Metadaten das Open-Data-Prinzip verankert. Bei den eigentlichen Inhalten wird die Respektierung der Public Domain eingefordert, aber ansonsten viel Spielraum gelassen. Erwünscht ist allerdings, dass Gedächtnisinstitutionen nicht nur gemeinfreie Inhalte für die freie Weiterverwendung durch Dritte bereitstellen, sondern auch alle übrigen Inhalte aus ihren Beständen, sofern keine urheberrechtliche oder andere rechtliche Gründe dagegen sprechen. Damit wird nämlich die Nutzung der Daten durch Verringerung der Transaktionskosten merklich erleichtert. Neben der Öffnung von Daten und Inhalten steht auch das Schaffen neuer Partizipationsformen im Vordergrund, die durch das Internet ermöglicht werden. Damit lässt sich OpenGLAM als logische Fortschreibung der Entwicklung verstehen, die mit dem Aufkommen des Internets (Web 1.0 und Web 2.0) und der zunehmenden Digitalisierung von Überlieferungsobjekten angestossen wurde.

Frédéric Noyer wird den im Rahmen des Schweizer Kulturhackathons verfolgten Ansatz vorstellen und mit Ergebnissen der Teilnehmerbefragungen der beiden ersten Schweizer Kulturhackathons aufwarten können. Diese umfassen Angaben zur Akzeptanz der OpenGLAM-Prinzipien und der Evaluierung ihres Nutzens für die Forschung in den Digital Humanities, zu den bisherigen Hackathon-Erfahrungen der Teilnehmenden, zu ihrer Rolle während des Anlasses, zu ihren Aktivitäten während und nach dem Hackathon sowie zur Zufriedenheit und zur Wahrnehmung der Wirksamkeit des Hackathons in Bezug auf verschiedene Ziele, wie Knowhow-Austausch, Networking, das Generieren von neuen Ideen, das Umsetzen von Projekten oder die Förderung von Open Data unter Gedächtnisinstitutionen.

Projekt “Visual Exploration of Vesalius' Fabrica”

Danilo Wanner, YAAY, Basel (unter Vorbehalt)

Danilo Wanner ist Designer und Mitglied von YAAY, eine Informationsdesign-Agentur aus Basel. Ihre Arbeit besteht darin, durch kreatives Design komplexe Informationen zu konsolidieren, umzustrukturieren und zu vereinfachen.

Das Team von YAAY schloss sich am Open Cultural Hackathon mit Radu Suciu zusammen, einem Spezialisten für frühneuzeitliche medizinische Publikationen. Das Team befasste sich mit dem Digitalisat des medizinischen Buchs “De humani corporis fabrica” des niederländischen Anatomisten Adreas Vesualis aus dem 16. Jahrhundert. Suciu hatte das Buch im Rahmen seiner Dissertation untersucht. [2]

Die Originalversion des Buchs liegt in der Universitätsbibliothek Basel, dem Veranstaltungsort des zweiten Schweizer Open Cultural Hackathons. Das Buch wurde von der Bibliothek digitalisiert und in einer PDF-Version online zur Verfügung gestellt. Im Gespräch mit Suciu wurde klar, dass es nicht möglich ist, aufgrund der PDF-Version den Wert und die Bedeutung des Buchs richtig einzuschätzen. Da das Buch unter strengen konservatorischen Vorschriften gelagert werden muss, ist es nur schwer zugänglich. Inhaltlich ist das Buch ausserdem nur mit grossem Fachwissen verständlich.

Das Ziel dieser Zusammenarbeit von Forscher und Designern war nun, die digitale Version des Buchs zu analysieren. Die Seiten des Buches wurden in verschiedene Kategorien – Illustrationen, Text und Kombination dieser beiden – eingeteilt. Aufgrund dieser Analyse wurden vier “Stories” erstellt und auf einem Farbbalken visualisiert. Basierend auf den Metadaten des Digitalisats wurde eine Infografik mit interessanten Fakten zum Buch erstellt.

Diese Art der Buchanalyse könnte auf alle Bücher mit einem seltenen und/oder wertvollen Inhalt angewendet werden, um Leserinnen und Lesern den Zugang zu vereinfachen. Die Analysemethode könnte ausserdem auf weitere Kunst- oder Wissenschaftsformen wie Ausstellungen oder Bilder übertragen werden.

Das Ergebnis des Projekts ist ein Prototyp eines Software-Programms, das an der Konferenz vorgestellt wird. Ausserdem wurden durch den Vergleich der Illustrationen interessante Entdeckungen zur Untersuchung

des Buchs gemacht. Die verschiedenen Resultate werden im Referat präsentiert.

Dodis goes Hackathon
Christiane Sibille & Sacha Zala, Dodis, Bern

Die Diplomatischen Dokumente der Schweiz (DDS) sind ein Projekt zur Edition zentraler Dokumente zur Geschichte der schweizerischen Aussenbeziehungen. Ihre Datenbank Dodis ermöglicht den freien Zugang zu einer grossen Anzahl von digitalisierten Dokumenten und liefert Informationen zu in- und ausländischen Personen und Körperschaften, die aussenpolitisch aktiv waren.

Für den ersten Swiss Cultural Hackathon stellten die DDS Daten zu den vorhandenen Dokumenten zur Verfügung. Ein Schwerpunkt lag hierbei auf geographischen Metadaten, die kurz vor dem Hackathon vollständig geolokalisiert wurden. Seit 2015 stehen diese Daten unter opendata.dodis.ch zur Verfügung.

Bei beiden Hackathons stiessen die zur Verfügung gestellten Daten auf grosses Interesse, wobei insbesondere die fertige Geolokalisation von den Teilnehmenden sehr geschätzt wurde.

Dies hat gezeigt, dass qualitativ hochwertig aufbereitete Forschungsdaten nicht nur für die Scientific Community relevant sind, sondern auch für die Weiterverarbeitung durch eine interessierte Öffentlichkeit attraktiv sind. Die Hackathons haben DDS darin bestärkt, das Prinzip der digitalen Offenheit, das auf den drei Säulen Open Access, Open Source und Open Data basiert fortzusetzen und weiter auszubauen.

Aufbauend auf Erfahrungen, die während des Hackathons gemacht wurden, haben die DDS daher auch ihr Engagement im Bereich Linked Open Data verstärkt.

Die Präsentation wird auf die Rolle des Hackathons für die Langzeitstrategie in Bezug auf Innovation und Kommunikation eines Forschungszentrums wie Dodis eingehen. Ausserdem wird kritisch auf die gemachten Erfahrungen im Rahmen der zwei Hackathons eingegangen.

VSJF-Flüchtlingsmigration zwischen 1898-1975 in der Schweiz
Maria-Elisabeth Züger, Archiv für Zeitgeschichte, Zürich

Unsere Gruppe bestand aus 6 Mitgliedern aus unterschiedlichen Bereichen. Es kamen Entwickler mit Datenspezialisten sowie Archivaren und dem Datenlieferant zusammen.

Die Daten sind ein Auszug aus der VSJF-Datenbank, welche durch das Archiv für Zeitgeschichte an der ETH Zürich (Datenlieferant) verwaltet wird. Die Datenbank beinhaltet Daten zu Flüchtlingen, welche bei dem Verband Schweizerischer Jüdischer Fürsorgen (VSJF) verzeichnet sind. Sie ist Resultat eines langjährigen Projektes in welchem Inhalte aus Akten des VSJF händisch übertragen wurden.

Auf Basis des VSJF-Datenauszuges entwickelten wir eine interaktive Visualisierung des Migrationsflusses zur und durch die Schweiz im Zeitraum von 1898-1975. Um den Fluss zu veranschaulichen nutzten wir eine Schnittstelle zu Google Maps. Die Visualisierung wird in einer HTML-Seite mittels JavaScript dargestellt. Auf einer Karte wird der Migrationsfluss von über 20,000 Flüchtlingen im zeitlichen Verlauf abgebildet.

Der Weg eines Flüchtlings beginnt beim Geburtsort und setzt fort mit dem Ort von dem aus die Einreise in die Schweiz stattfand (sofern bekannt). Dann folgt eine Reihe an Aufenthalten in der Schweiz. Schlussendlich verlässt der Flüchtling die Schweiz wieder von einem Ort in der Schweiz zu einem Bestimmungsort ausserhalb des Landes.

Eine wichtige Aufgabe während des Hackathons war es die Daten für die Visualisierung vorzubereiten. Der Fluchtweg musste aus den vorhandenen Daten extrahiert und für die Schnittstelle aufbereitet werden. Hierfür verwendeten wir Google Refine (jetzt bekannt unter OpenRefine) und reguläre Ausdrücke zur Textmustererkennung um die Daten in mehreren Schritten automatisiert zu strukturieren. Lager und Aufenthaltsorte wurden geocodiert und nach Typen kategorisiert. Die Kategorien sind auf der Karte mittels Farben dargestellt. Zudem wurden relevante historische Ereignisse recherchiert und als Hintergrundinformation in die Visualisierung eingebettet.

Nach dem Hackathon erhielten wir positives Feedback des Archivs für Zeitgeschichte und eine Diskussion über Nutzen und Möglichkeiten einer derartigen Visualisierung fand statt.

Moderiert wird das Panel von **Jan Baumann**. Er ist Mitarbeiter von infoclio.ch und hat die beiden ersten Ausgaben des Schweizer Kulturhackathons mitorganisiert.

In der Diskussion wird es unter anderem um folgende Fragen gehen:

- Welchen Nutzen haben Hackathons für die Forschung in den Digital Humanities? Welchen Nutzen bieten Hackathons in methodologischer Hinsicht für die Forschung?
- Welche Bedeutung hat die Open-Data-Politik für die digitale Nachhaltigkeit? Kann die langfristige Erhaltung von digitalen

Daten durch die freie Zugänglichkeit besser gesichert werden, als wenn der Zugang beschränkt bleibt?
- Wie gelingt die Integration der unterschiedlichen Stackholder (Programmierer, Vertreter der Daten-Lieferanten, Forschende, Wikipedianer, Designer) währende eines Hackathons?
- Wie steht es um die Nachhaltigkeit der am Hackathon entwickelten Projekte? Werden Projekte nach dem Hackathon weiterentwickelt? Gibt es Best Practices für eine nachhaltige Weiterentwicklung der Projekte?

Fußnoten

1. See http://openglam.org/principles/
2. Radu Suciu, André du Laurens - Discours des maladies mélancoliques (1594), Paris, Klincksieck, 2012.

Bibliographie

Briscoe, Gerard / Mulligan, Catherine (2014): „Digital Innovation: The Hackathon Phenomenon“, in: *Working Papers of The Sustainable Society Network* https://qmro.qmul.ac.uk/xmlui/handle/123456789/11418 [letzter Zugriff 1. Dezember 2016].

Decker, Adrienne / Eiselt, Kurt / Voll, Kimberly (2015): „Understanding and Improving the Culture of Hackathons: Think Global Hack Local“, in: *Presentations and other scholarship*, 30. September 2015. http://scholarworks.rit.edu/other/847 [letzter Zugriff 1. Dezember 2016].

Groen, Derek / Calderhead, Ben (2015): „Science hackathons for developing interdisciplinary research and collaborations“, in: *eLife* 10.7554/eLife.09944.

Johnson, Peter / Robinson, Pamela (2014): „Civic Hackathons: Innovation, Procurement, or Civic Engagement?“, in: *Review of Policy Research* 31 (4): 349–357 10.1111/ropr.12074.

Komssi, Marko / Pichlis, Danielle / Raatikainen, Mikko / Kindström, Klas / Järvinen, Janne (2015): „What are Hackathons for?“, in: *IEEE Software* 32 (5): 60–67.

Melissa, Gregg (2015): „FCJ-186 Hack for good: Speculative labour, app development and the burden of austerity“, in: *The Fibreculture Journal* 25 http://twentyfive.fibreculturejournal.org/fcj-186-hack-for-good-speculative-labour-app-development-and-the-burden-of-austerity [letzter Zugriff 01. Dezember 2016].

Prahalad, C.K. / Ramaswamy, Venkat (2004): „Co-creation experiences: The next practice in value creation“, in: *Journal of Interactive Marketing* 18 (3): 5–14 10.1002/dir.20015.

Pui Ying To, Jacqueline (2016): *Understanding the Potential of Public Engagement: Hackathons and Jams*. Master Thesis, OCAD University, Toronto http://openresearch.ocadu.ca/666/1/To_Jacqueline_2016_SFIN_MRP_withRevisions.pdf [letzter Zugriff 1. Dezember 2016].

Nachhaltige Entwicklung digitaler Ressourcen und Werkzeuge für wenig erforschte historische Sprachen

Feige, Tillmann
tillmann.feige@uni-hamburg.de
Universität Hamburg, Deutschland

González, Alicia
alicia.gonzalez@uni-hamburg.de
Universität Hamburg, Deutschland

Prager, Christian
cprager@gmx.net
NRW Akademie der Wissenschaften und der Künste

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

Werwick, Heiko
heiko.werwick@web.de
Universität Jena, Deutschland

Das Panel wird durch vier Kurzvorträge in Probleme der Nutzung digitaler Werkzeuge für nicht-indoeuropäische Sprachen einführen. Die Vorträge basieren auf Erfahrungen aus langfristig angelegten Projekten und haben jeweils individuelle Lösungen für die spezifischen Anforderungen gefunden, die hauptsächlich durch die Sprache formuliert werden.

Allen Projekten gemein sind Probleme der Nutzung von digitalen Werkzeugen, die insbesondere bei der Verwendung von Quellen historischer oder wenig erforschter Sprachen auftreten.

Aktuelle Content Management Systeme und Annotationstools wurden selten im Hinblick auf Anforderungen aus Orchideenfächern entwickelt. Dies betrifft beispielsweise einige Sprachen mit nichtkonkatenativer Morphologie oder komplexen Schriftsysteme. Daher müssen für erwähnte Sprachen entweder existierende Anwendungen angepasst oder erschaffen werden.

Bei der Adaption können während der Modellierung wichtige Eigenschaften nicht berücksichtigt werden oder bleiben nur als Kommentar erhalten, was eine weitere maschinelle Bearbeitung erschwert. Bezüglich der Datenkodierung ergibt sich das Problem der Ineffizienz. So wurden morphologische Tagsets primär für die indo-europäische Sprachfamilie entwickelt. Für eine tiefe linguistische Annotation müssen aber diese Standards beispielweise für einige semitische Sprachen angepasst werden.

Nicht selten ist die Alternative die Eigenentwicklung projektbezogener Lösungen, die aber aufgrund der Anforderungen mit eigenen Datenformaten arbeiten, und so nicht mehr den geltenden Standards folgen und den Austausch erschweren. Hinzu kommt der immense Zeit- und Ressourcenaufwand bei der Implementierung.

Allerdings sind gerade im deutschsprachigen Raum viele langfristige Projekte auf digitale Tools angewiesen.

Durch eine Vernetzung solcher Projekte können gemeinsame Anforderungen an, und Begrenzungen von aktuellen Lösungen besprochen und Initiativen zur Entwicklung digitaler Tools und Ressourcen koordiniert werden. Daher ist das Ziel dieses Panels eine erste Zusammenführung langfristig ausgerichteter Projekte im deutschen Sprachraum, die mit historischen nicht-indo-europäischen Sprachen im digitalen Kontext arbeiten. Dabei sollen die Probleme der Nachhaltigkeit entwickelter Werkzeuge und Ressourcen, sowie der bearbeiteten Daten besprochen werden. Anschließend werden die vielfältigen Herangehensweisen mit einem Fokus auf drei große Punkte diskutiert:

- Nachhaltigkeit von Repositorien
  - Welche Frameworks werden für welche Datentypen benötigt?
  - Wie können Informationen über unpräzise Daten gespeichert werden?
  - Wie gehen verfügbare Systeme mit Multilingualität um?
- Nachhaltigkeit von (Annotations-)Werkzeugen
- Analyse historischer Daten impliziert die Annotation von Textmaterialien in Sprachen, die aus verschiedenen Gründen zu Problemen führen.
  - Welche Annotationstools können genutzt werden? Mit welchen Limitierungen?
  - Was bedeutet es, ein neues Tool zu entwerfen?
  - Häufige Anforderungen durch strukturell komplexe Sprachen: Multilevel-Annotation, Textkorrektur während der Annotationsphase, Multilevel-Segmentierung
- Nachhaltigkeit des annotierten Materials (Standards)
- Während der Standard TEI-XML als Schnittstellenformat sehr nützlich ist, ergeben sich dennoch Probleme wie:
  - für interne Verarbeitung kann dessen Verwendung hinderlich sein. Daher müssen projekt-spezifische Lösungen mit standardisiertem Export entwickelt werden.
  - Können diese Daten von Dritten in TEI-XML verarbeitet werden?
  - Welche anderen Formate können genutzt werden (z.B. JSON)?
  - Sind existierende Tagset-Formate ausreichend spezifiziert, um auch nicht-europäische Sprachen taggen zu können?

Herausforderungen in der Nutzung vorhandener Tools für arabische Daten

Alicia González, Tillmann Feige
Universität Hamburg
ERC- Projekt COBHUNI (https://www.cobhuni.uni-hamburg.de/)
Email: alicia.gonzalez@uni-hamburg.de; tillmann.feige@uni-hamburg.de

Wir beschreiben den Ansatz, einen Korpus der neben modernem auch klassisches Arabisch (siehe Romanov, 2016) enthält, mit

computerlinguistischen und semantischen Verfahren analysierbar zu machen. Wir setzen auf bereits vorhandene Software für die Hauptpunkte Annotation und Analyse. Dazu wurde ein Pflichtenheft erstellt, dass mit vorhandenen Tools abgeglichen wurde.

Da wir mit arabischen Daten arbeiten, ist eine große Herausforderung die Schrift. Es ist eine linksläufige verbundene Schrift, die durch Konsonanten und lange Vokale repräsentiert wird. Kurze Vokale sind Diakritika, die optional gesetzt werden und gerade bei Referenzen auf religiöse Quellen im Textkorpus vorkommen. Dabei ist vollständige UTF-8 Unterstützung und die saubere Darstellung der Schrift unabdingbar. Dies reduziert die Auswahl erheblich. Hinzu kommt, dass wir auf flexible Import- und Exportmöglichkeiten angewiesen sind. Ähnliche Probleme führen Peralta und Verkinderen auf (Peralta / Verkinderen 2016). Durch unsere Herangehensweise gibt es weitere Einschränkungen wie Mehrebenen-, Multitoken- aber auch Subtoken-Annotation.

Die Auswahl für die semantische Annotation fiel auf WebAnno, dass durch sein spezielles Datenmodell die erforderliche Datenaufbereitung und Kontrolle gestattet.

Als Visualisierungstool haben wir ANNIS ausgewählt, dass ebenfalls Arabisch unterstützt, einen konfigurierbaren Converter mitbringt und Mehrebenenkorpora erlaubt, so dass auch hier die Hauptkriterien erfüllt wurden. Zusätzlich lassen sich potentielle Probleme in der Darstellung durch eine anpassbare HTML-Visualisierung umgehen. Durch Zusammenarbeit mit den Entwicklern beider Programme wurde die Unterstützung für Arabisch stetig ausgebaut.

Im Beitrag werden wir die einzelnen Punkte erläutern und darstellen, warum wir uns für die angeführten Programme und gegen eine Eigenentwicklung entschieden haben, sowie welche Implikationen diese Entscheidung für die Nachhaltigkeit des Projekts, der Daten und der genutzten Tools hat.

Tiefe Mehrebenen-Annotation für semitische Sprachen: der Fall von Ge'ez

Cristina Vertan
Universität Hamburg
ERC-Projekt TraCES (https://www.traces.uni-hamburg.de/)
Email: cristina.vertan@uni-hamburg.de

Das südsemitische Gəʿəz ist die Sprache des Königreichs Aksum in der heutigen nordäthiopischen Provinz Tigray, von wo aus die im 4. Jahrhundert beginnende Christianisierung Äthiopiens ihren Anfang nahm. Die in der Folge entstehende reiche Literatur ist in großem Umfang geprägt von Übersetzungen, was durch grammatische Interferenzphänomene reflektiert wird. Das Altäthiopische hat aus einer südsemitischen Schrift ein eigenes Silbenalphabet entwickelt, das bis heute in mehreren modernen Sprachen Äthiopiens und Eritreas Verwendung findet. Innerhalb der semitischen Sprachen fällt es durch die verwendete Rechtsläufigkeit auf; außerdem werden die Vokale vollständig geschrieben. Beides unterscheidet das Gəʿəz von verwandten Sprachen wie Altsüdarabisch, Arabisch, Hebräisch und Syro-Aramäisch Mit den genannten eng verwandten semitischen Sprachen teilt das Altäthiopische die nichtkonkatenative Morphologie. Durch das äthiopische Silbenalphabet sind Morphemgrenzen in der Schrift nicht darstellbar, so dass beispielsweise ein einzelner Vokal als Bestandteil einer Silbe eine eigenständige Wortart darstellt und tokenisiert werden muss.

Die Komplexität des Annotationstools wird sehr vielfältige linguistische Anfragen und detaillierte Analysen der Sprache ermöglichen, aber auch eine vollautomatische Annotation verhindern. Ein alle morphologischen Merkmale abdeckendes Vektorraum-Modell (das für maschinelle Lernverfahren benutzt werden muss) wäre zu groß. Vorstellbar ist lediglich eine flache automatische Annotation (z. B. der Wortarten); jedoch wird auch für eine solche zunächst eine relativ große Menge an Trainingsdaten benötigt. Daher ist die Entwicklung eines Werkzeugs für die manuelle Annotation ein obligatorischer Schritt.

Die Besonderheit der entwickelten Lösung (Vertan/Ellwardt/Hummel 2016) sind:

- automatische Transkription
- manuelle Korrektur der Transkription während des Annotationsprozesses
- semi-automatische Verfahren: automatische Verläufe werden farbig markiert und sind automatisch zur manuellen Korrektur hinterlegt
- Mehrebenenannotation: Linguistik, Edition, Textstruktur
- Anpassungen an unterschiedliche Schriftsysteme und Transkriptionsregeln

Nutzungs- und Nachhaltigkeitsstategien im Projekt "Textdatenbank und Wörterbuch des Klassischen Maya"

Christian M. Prager
NRW Akademie der Wissenschaften und der Künste
http://mayawoerterbuch.de/
Email:

Die Mayaschrift ist das einzig lesbare Schriftsystem der vorspanischen Amerikas. Die über 10.000 Texte sind in einer logographisch-syllabischen Hieroglyphenschrift verfasst und von den rund 800 Zeichen sind erst 60% sicher entziffert. Die Texte enthalten taggenaue Kalenderangaben, die es uns ermöglichen die rund 2000jährige Sprach- und Schriftgeschichte genau zu dokumentieren. Das Projekt (Prager 2015) wird sämtliche Inschriften einschließlich Metadaten in einer Datenbank einzupflegen und darauf basierend ein digitales Wörterbuch des Klassischen Maya zu kompilieren. Herausforderung dabei ist, dass die Schrift noch nicht vollständig entziffert ist und bei der Modellierung zu berücksichtigen ist. Unser Projekt verfolgt den Ansatz, wonach die Bedeutung von Wörtern ihre Verwendung ist - Texte nehmen Bezug auf den Textträger und den Verwendungskontext und nur die exakte Dokumentation sogenannter nicht-textueller Informationen erlaubt es, textuelle und nicht-textuelle Informationsbereiche zueinander in Beziehung zu setzen und bei der Entzifferung von Zeichen und Textstellen zu berücksichtigen. Zum Zweck der Nachhaltigkeit und Nachnutzung greift das Projekt bei der Beschreibung der Artefakte und der relevanten objektgeschichtlichen Ereignisse auf CIDOC CRM zurück, das eine erweiterbare Ontologie für Begriffe und Informationen im Bereich des kulturellen Erbes anbietet. Das entstandene Anwendungsprofil wird durch Elemente aus weiteren Standards und Schemata angereichert und wird damit auch für vergleichbare Projekt nachnutzbar. Die Schemata und erstellten Metadaten werden in einer Linked (Open) Data-Struktur (LOD) abgebildet. Durch die Repräsentation im XML-Format, sowie die Nutzung von HTTP-URIs wird eine einfache Austauschbarkeit und Zitierbarkeit der Daten ermöglicht. Durch diese Umsetzung können Objektmetadaten getrennt vom erfassten Text gespeichert werden und durch die Verwendung der HTTP-URI verlinkt werden. Die Nachnutzung bereits bestehender und fachlich anerkannter Terme trägt darüberhinaus auch zu einer hohen Interoperabilität mit anderen Datenbeständen und Informationssystemen bei. Das ausgestaltete Schema hat eine ontologisch-vernetzte Struktur, die komplexe Beziehungen und Zusammenhänge abbildet.

Interdisziplinäre Digitale Zusammenarbeit für seltene Sprachen und Kulturen

- Eine Fallstudie über jiddische Texte aus der frühen Neuzeit -

Walther v. Hahn (Universität Hamburg), Berndt Strobach (Wolffenbüttel)
Email: vhahn@informatik.uni-hamburg. de, berndt.strobach@freenet.de>

In den Geisteswissenschaften werden häufig die fachlichen Interpretationen und die sprachlichen Erklärungen von verschiedenen Gruppen mit unterschiedlicher Kompetenz bearbeitet. Gute Beispiele sind Studien zu Texten aus semitischen Sprachen, wobei, speziell bei historischen Dokumenten die historische oder geistes- und sozialgeschichtliche Würdigung von Forschern verfasst werden muss, die des Hebräischen, Arabischen, Aramäischen etc. nicht mächtig sind, die sprachwissenschaftlichen Forscher dagegen bei der Interpretation gelegentlich weniger engagiert bleiben. Extremfälle wie Studien über das Sephardische in Spanien (Ladino, Djudezmo) machen etwa solide Kenntnisse zumindest des Spanischen, Hebräischen, Türkischen, Griechischen und Italienischen zur Voraussetzung für seriöse hermeneutische Forschungsergebnisse.Wir berichten über Studien zu jiddischen Texten aus dem Wolffenbüttel des 18. Jahrhunderts, in denen die Rolle der "Hofjuden" und ihres kultur- und sozialgeschichtlichen Hintergrundes diskutiert wird.

Die Herausforderung einer interdisziplinären Zusammenarbeit zwischen Historikern, Sprachwissenschaftlern und Informatikern besteht darin,

1. die Lesbarkeit der Originalquellen für alle Gruppenmitglieder sicher zu stellen (Invertierte Transkriptionen, Vokalisierung, Visualisierung), sowie

2. in der Gruppe eine gemeinsame Behandlung von Vagheit, Unsicherheit und Unbekanntem zu definieren, so dass die Unklarheiten in den einzelnen Forschungsstufen erhalten und im Endergebis sichtbar bleiben (Vagheits-Annotationen und vage Inferenzen). Heute werden derartige Unsicherheiten meist bereits in den Annotationen unterschlagen (von Hahn, 2016).

Bibliographie

Hahn, Walther von (2016): „Humanities meet Computer Science – Digital Humanities between Expectations and Reality“, zu erscheinen in: von Hahn, Walter / Papadima, Liviu / Vertan, Cristina (eds.): *Humanities2020, New Trends in Education and Research*. Bukarest: University of Bucharest Publishing House.

Peralta, José Haro / Verkinderen, Peter (2016): „‚Find for me!‘: Building a Context-Based Search Tool Using Python“, in: Muhanna, Elias (ed.): *The Digital Humanities and Islamic & Middle East Studies*. Berlin: Walter de Gruyter GmbH 199–231.

Prager, Christian M. (2015): „Das Textdatenbank- und Wörterbuchprojekt des Klassischen Maya: Möglichkeiten und Herausforderungen digitaler Epigraphik“, in: Neuroth, Heike / Rapp, Andrea / Söring, Sibylle (eds.): *TextGrid: Von der Community - für die Community: Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften*. Glückstadt: Werner Holsbusch 105–124 https://www.academia.edu/17957108/Das_Textdatenbank-_und_W%C3%B6rterbuchprojekt_des_Klassischen_Maya_M%C3%B6glichkeiten_und_Herausforderungen_digitaler_Epigraphik .

Romanov, Maxim (2016): *Creating Frequency-Based Readers for Classical Arabic* http://maximromanov.github.io/2016/05-30.html [letzter Zugriff 1. Dezember 2016].

Vertan, Cristina / Ellwardt, Andreas / Hummerl, Susanne (2016): „Ein Mehrebenen-Tagging-Modell für die Annotation altäthiopischer Texte“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* http://www.dhd2016.de/abstracts/vorträge-061.html .

Virtuelle Forschungsplattformen im Vergleich: MONK, Textgrid, Transcribo und Transkribus

Piotrowski, Michael
piotrowski@ieg-mainz.de
Université de Lausanne

Schomaker, Lambert
l.r.b.schomaker@rug.nl
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Horstmann, Wolfram
horstmann@sub.uni-goettingen.de
Universität Trier

Burch, Thomas
burch@uni-trier.de
Staatsarchiv des Kantons Zürich

Hodel, Tobias
tobias.hodel@hist.uzh.ch
Leibniz-Institut für europäische Geschichte Mainz

Eine zentrale Forderung zur Unterstützung digitaler Editionen ist das Anbieten virtueller Umgebungen (Interfaces, Software) zur Produktion, aber auch zum Management digitaler Daten (BMBF 2013). In den letzten Jahren wurden aufgrund dieser durch FachwissenschaftlerInnen getragenen Nachfrage mehrere Plattformen und Softwareangebote/Infrastrukturen geschaffen, die Prozesse der digitalen Datenerstellung von der Aufnahme von Informationen (Metadaten, Transkriptionen) über die Auswertung und Anreicherung bis zur Publikation unterstützen (DARIAH-DE (Hg.), 2015) und nachhaltig betrieben werden sollen. Unterschiedliche Konzepte und angebotene Abläufe sowie integrierte Hilfsmittel stehen für eine je eigene Profilierung der Plattformen. Merkmale der Angebote, insbesondere Leistungsfähigkeit, unterstützte Prozesse und Ausrichtungen unterscheiden sich zwangsläufig. Im Panel werden aus

diesem Grund wichtige und häufig eingesetzte Plattformen in ihrem Leistungsumfang verglichen und einander gegenübergestellt. Im Sinne geisteswissenschaftlicher software studies (Andrews, 2016) müssen die Plattformen nicht nur aus pragmatischen Gründen gegeneinander abgewogen werden sondern auch, um in den angebotenen Prozessen angelegte Praktiken auf ihre Logik und dadurch entstehende Folgen zu untersuchen (Drucker, 2013). Anhand eines klar umrissenen Fragebogens präsentieren Monk, Textgrid, Transcribo und Transkribus Arbeitsabläufe, Services und Vernetzungsmöglichkeiten. Damit wird Interessierten in einem Panel aus erster Hand ein Vergleich wichtiger, produktiv nutzbarer Angebote geliefert.

Das Panel wird moderiert von Michael Piotrowski (IEG Mainz).

Folgende Frage- und Themenschwerpunkte werden schriftlich und in kurzen Präsentationen dargeboten:

- Idealtypischer/Schematisierter Ablauf für den Gebrauch der Plattform
- Zeitliche Anforderungen, um ein Projekt aufzusetzen/ein Dokument zu verarbeiten; zu exportieren
- Herstellung von Transkriptionen
- Bild-Text-Verknüpfung
- Text-Markup
- Ausgabemöglichkeiten (für Edition und/oder Transkription)
- Vernetzungsmöglichkeiten (Wörterbücher, externe Ressourcen, Ontologien)
- Datei-/Bildverwaltung
- Projektverwaltung
- Auswertungs-/Abfrageoptionen
- Automatisierungen
- Crowdsourcing/Optionen zum Einbezug von Laien oder Externen
- Nachhaltigkeit der Plattform/der enthaltenen Daten
- Updates bis 2018

Monk (presented by Lambert Schomaker, Rijksuniversiteit Groningen)

The Monk system is a trainable search engine for handwritten material. For the humanities, it may serve as a method for getting keyword access to scanned pages at the earliest stages after a document digitisation. For pattern recognition research, it is an observatory for complicated visual material and its human-provided labels (e.g., word or character labels). The system act as an e-Science service that is continuously available.

An internal image and metadata format is used, which can be exported to, e.g., PAGE xml. Provisional transcriptions can be retrieved as flat text. Indices can be exported upon request.

The system makes a distinction between four different forms of annotation: page (scan) descriptors, typically page titles, page regions of interest (tags for visual objects), transcription of segmented lines, and finally, word labeling. The system could export in TEI, however, within the OCR community, there is a preference for layout-centric description languages, as opposed to editorial descriptions. In practice, both TEI and PAGE are used, as well as other formalisms that allow to provide metadata to polygonal image sections.

In order to proceed data in Monk, scans are uploaded via sftp or mailed hard disks. The collection is then judged on the required preprocessing steps (multicolumn, contrast enhancement, line segmentations), and 'ingested'. Within one or two days users can start to label words. The system performs data mining on the collection and presents hit lists for words which can be labeled further, and so on. Static indices and provisional transcriptions are updated nightly.

At the moment 400 documents from different periods and handwriting styles are being processed. The Monk system is one of the first 24/7 machine learning systems. The system detects where compute resources should be directed, on the basis of observed user activities and interests.

The Monk system is part of the large multi-petabyte Target platform of the university of Groningen, in collaboration with astronomy, genomics and the IBM company.

TextGrid (präsentiert durch Wolfram Horstmann, Niedersächsische Staats- und Universitätsbibliothek Göttingen)

Hintergrund

Die Entwicklung von TextGrid, einer Virtuellen Forschungsumgebung für die Geistes- und Kulturwissenschaften, wurde durch die zunehmende Nachfrage aus den Fachwissenschaften nach digitalen Werkzeugen v.a. des philologischen Edierens und kollaborativen Arbeitens angestoßen. Das Bundesministerium für Bildung und Forschung (BMBF) hat TextGrid als Verbundprojekt mit über zehn institutionellen und universitären Partnern zwischen 2006 und 2015 gefördert.

Die Software steht mittlerweile in einer stabilen Version 3.0 zum kostenfreien Download bereit. Software, Archiv und damit das gesamte Angebot werden in Zusammenarbeit mit AnwenderInnen, FachwissenschaftlerInnen

und Fachgesellschaften und in Kooperation mit DARIAH-DE - Digital Research Infrastructure for the Arts and Humanities weiter entwickelt und dauerhaft betrieben.

Zielpublikum

FachwissenschaftlerInnen, die mit TextGrid Forschungsprojekte wie z.B. digitale Editionen erarbeiten

EntwicklerInnen, die TextGrid-Tools und Services für eigene Vorhaben anpassen oder externe Services und Tools in TextGrid integrieren

Forschungsprojekte und -institutionen, die Daten in TextGrid archivieren und für Dritte zugänglich und nutzbar machen (Repository)

Form des Einsatzes

Die virtuelle Forschungsumgebung (VFU) TextGrid unterstützt digital arbeitende GeisteswissenschaftlerInnen im gesamten Forschungsprozess – insbesondere beim Erstellen digitaler Editionen.

Sie besteht aus drei Kernbereichen:

- Die Software **TextGrid Laboratory** stellt den Einstiegspunkt in die VFU dar und bietet unterschiedliche Open-Source-Werkzeuge und -Services für den gesamten Forschungsprozess zur Verfügung, z. B. einen Text-Bild-Link Editor für die Verknüpfung von Digitalisaten und Transkriptionen
- Im **TextGrid Repository**, einem Langzeitarchiv für geisteswissenschaftliche Forschungsdaten, können XML / TEI-kodierte Texte, Bilder und Datenbanken sicher gespeichert, publiziert und durchsucht werden.
- Die beständig wachsende **TextGrid Community** trifft sich bei regelmäßigen Nutzertreffen zu themen- bzw. anwendungsspezifischen Workshops, die nicht zuletzt auch den Austausch zwischen digitalen Forschungs- vorhaben aus den Geisteswissenschaften befördern.

Eine Stärke

TextGrid unterstützt den gesamten wissenschaftlichen Arbeitsprozess im Rahmen der Erstellung digitaler Editionen vom Ingest des Ausgangsmaterials (Text- und/ oder Bilddatei- en / Faksimiles) über die Anreicherung und Auszeichnung der Daten (Annotationen, Verknüpfungen) bis zur Veröffentlichung (Portal, Print) und nachhaltigen Archivierung (Repository) und wird stetig basierend auf konkreten fachwissenschaftlichen Anforderungen weiterentwickelt.

Eine Schwäche

Technisch setzt TextGrid auf dem Eclipse-Framework auf, aus heutiger Sicht, wären webbasierte Tools wünschenswerter. Zugleich verdeutlicht dies, dass Softwareentwicklungen permanente Weiterentwicklung benötigen, um sich neuen technologischen aber auch sich wandelnden User-Requirements stellen zu können.

Transcribo (präsentiert durch Thomas Burch, Universität Trier)

Transcribo wird in enger Zusammenarbeit von Philologen und Informatikern der Kooperationspartner entwickelt. Die grafische Nutzeroberfläche ist um das digitale Faksimile, also in der Regel den gescannten Überlieferungsträger, zentriert. Beliebig große Einheiten (z.B. Wörter, Zeilen oder Absätze) können mittels eines Rechteck- oder Polygonwerkzeugs markiert, transkribiert und annotiert werden. Dabei wird jede Bilddatei doppelt dargeboten: links liegt das Original zur Ansicht, die rechte Version dient als Arbeitsunterlage, hier wird der transkribierte Text topografisch exakt über das leicht ausgegraute Faksimile gelegt. Wo die räumliche Anordnung nicht der textuellen Wortreihenfolge entspricht, können Wörter in der grafischen Oberfläche zu Sequenzen zusammengefasst und so die semantischen Zusammenhänge im Transkript protokolliert werden. Ein zentrales Merkmal des Programms liegt außerdem in der Möglichkeit, in jeder erfassten Einheit textgenetische und editionsphilologisch relevante Phänomene zu kennzeichnen und mit Annotationen zu versehen. Dabei kommt ein Kontextmenü mit einer projektspezifischen Auswahl zum Einsatz. Diese umfasst bisher unterschiedliche Varianten von Korrekturen (wie etwa Sofortkorrekturen, Spätkorrekturen mit ein-, zwei- oder mehrfacher Durchstreichung und Überschreibung), die Kennzeichnung von Hervorhebungen sowie von unsicheren Lesungen oder nicht identifizierten Graphen. Diese Auswahl ist jedoch beliebig erweiterbar und wird über den gesamten Projektverlauf hinweg an die Erfordernisse der Textgrundlage angepasst.

Transkribus (präsentiert durch Tobias Hodel, Staatsarchiv Zürich)

Hintergrund

Transkribus ist eine Plattform, die zur automatisierten Erkennung und Annotierung von Texten dient. Sie leistet einerseits eine Verlinkung zwischen Text und Bild (auf Block, Zeilen und Wortebene), produziert andererseits standardisierte Exportformate (XML nach TEI-Standard, PDF, aber auch METS für die Integration in Repositorien). Damit steht eine vollausgerüstete Softwaresuite zur Verfügung, die von der Segmentierung über die Erkennung, Transkription und Edition bis zur Ausgabe alle

Schritte in der Herstellung hochwertiger Daten unterstützt.

Die im Projekt READ weiterentwickelte Software vereint somit praxisnah die Bedürfnisse von GeisteswissenschaftlerInnen und Aufbewahrungsinstitutionen mit den technischen Möglichkeiten und Angeboten, die momentan im Bereich der Informatik und Computerlinguistik ermöglicht werden.

Die Software steht in einer stabilen Version zum kostenfreien Download bereit. Das Projekt READ wird unterstützt durch das Horizon 2020 Forschungs- und Innovationsprogramm der Europäischen Union.

Zielpublikum

Aufbewahrungsinstitutionen, die eigene Bestände und Dokumente aufbereiten und zur Verfügung stellen wollen

Geisteswissenschaftlerinnen, die eigene Transkriptionen und Editionen in Transkribus erstellen wollen oder mit darin aufbereiteten Daten arbeiten

Interessierte Laien, die sich an Crowdsourcing-Initiativen beteiligen wollen

ComputerwissenschaftlerInnen, die mit den gewonnenen Daten arbeiten und eigene Algorithmen entwickeln oder verbessern wollen

Form des Einsatzes

Auf Transkribus werden Bilddateien hochgeladen, mit Layoutverlinkungen und Transkriptionen sowie Annotationen versehen. Unterstützt werden die Vorgänge durch Automatisierungsvorgänge im Bereich der Layouterkennung und der Transkription. Der Export der gewonnen Daten ist in unterschiedlichen Formaten möglich. Zusätzlich werden Module zum Crowdsourcing und zukünftig für e-Learning und Analyse mit Smartphone bereitgestellt.

Eine Stärke

Transkribus nutzt neueste Automatisierungsprozesse (u.a. mit rekursiven neuronalen Netzen) somit werden bestmögliche Resultate in Aussicht gestellt.

Eine Schwäche

Transkribus ist eine Expertensoftware und benötigt entsprechende Einarbeitungszeit, um die Dokumente effizient und zielgerichtet zu bearbeiten.

Bibliographie

Andrews, Tara (2015): „Software and Scholarship – Editorial“, in: *Interdisciplinary Science Reviews* 40: 342–348 10.1080/03080188.2016.1165456.

BMBF (Bundesministerium für Bildung und Forschung) (eds.) (2013): *Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften* https://www.bmbf.de/pub/forschungsinfrastrukturen_geistes_und_sozialwissenschaften

DARIAH-DE (ed.) (2015): *Handbuch Digital Humanities: Anwendungen, Forschungsdaten und Projekte* http://handbuch.io/w/DH-Handbuch .

Drucker, Johanna (2013): „Performative Materiality and Theoretical Approaches to Interface“, in: *DHQ: Digital Humanities Quarterly* 7 (1) http://digitalhumanities.org:8081/dhq/vol/7/1/000143/000143.html .

Schomaker, Lambert (2016): „Design considerations for a large-scale image-based text search engine in historical manuscript collections“, in: *Information Technology* 58 (2): 80–88 10.1515/itit-2015-0049.

Virtuelle Forschungsumgebung für objekt- und raumbezogene Forschung

Kuroczyński, Piotr
piotr.kuroczynski@herder-institut.de
Herder-Institut für historische Ostmitteleuropaforschung – Institut der Leibniz-Gemeinschaft, Deutschland

Stanicka-Brzezicka, Ksenia
ksenia.stanicka@herder-institut.de
Herder-Institut für historische Ostmitteleuropaforschung – Institut der Leibniz-Gemeinschaft, Deutschland

Fichtl, Barbara
barbara.fichtl@herder-institut.de
Herder-Institut für historische Ostmitteleuropaforschung – Institut der Leibniz-Gemeinschaft, Deutschland

Köhler, Werner
w.koehler@fotomarburg.de
Deutsches Dokumentationszentrum für Kunstgeschichte - Bildarchiv Foto Marburg, Deutschland

Brahaj, Armand
Armand.Brahaj@fiz-karlsruhe.de
FIZ Karlsruhe - Leibniz-Institut für Informationsinfrastruktur, Deutschland

Fichtner, Mark
m.fichtner@gnm.de
Germanisches Nationalmuseum Nürnberg, Deutschland

Die Digitalisierung der Gesellschaft hat längst alle Sparten unseres Lebens erfasst. Die computer-gestützte Forschung in den Geisteswissenschaften, allen voran die Computerlinguistik, blickt bereits auf eine über fünfzigjährige Tradition zurück. Mit dem informationstechnologischen Fortschritt der letzten drei Dekaden verfügt die zeitgenössische Wissenschaft über ein reichhaltiges, teils unüberschaubares Arsenal an digitalen Forschungswerkzeugen und Applikationen, Dokumentationsstandards, Datenformaten, etc. Gleichzeitig führen die neuen Informations- und Kommunikationstechnologien zu einem nie zuvor beobachteten Wachstum von Daten und Wissen (Abb. 1).

Abb. 1: Gesamtmenge an generierten Daten in den vergangenen Jahren (Quelle: http://edition.cnn.com/2014/11/04/tech/gallery/big-data-techonomics-graphs/ *)*

Diese Entwicklung stellt die Informationsgesellschaft vor neue Herausforderungen. Eine in letzten Jahren an Bedeutung gewinnende Vorgehensweise stellt die Strukturierung und Vernetzung der Forschungsdaten in einem mensch- und maschinenlesbaren Format. Einen entscheidenden Anteil an diesem Prozess nimmt die Idee von Semantic Web (Web 3.0) für sich in Anspruch (Berners-Lee / Hendler / Lassila, 2001). Mit der *Öffnung der Datensilos* und Verknüpfung der Daten geht die Idee einer semantischen Datenmodellierung und Disambiguierung der Forschungsdaten einher, die in ein weltweites Netzwerk miteinander in Verbindung stehenden Information (Linked Data) mündet. Diesen Ansatz folgend versuchen zurzeit viele Disziplinen ihre fachspezifischen Fragestellungen mit Hilfe einer sprachlich gefassten und formal geordneten Darstellung einer Menge von Begrifflichkeiten (Entitäten) und der zwischen ihnen bestehenden Beziehungen zu repräsentieren (Referenz- und Applikationsontologien). Diese konzeptionellen mensch- und maschinenlesbaren Wissensrepräsentationen können infolge einer Implementierung innerhalb einer Web Ontology Language (OWL) die generierten Forschungsdaten im RDF-Format von Linked Data vorhalten (Graphdatenbank). Für die Disambiguierung der digital vernetzen Datensätze stellt die Entwicklung und Zurverfügungstellung von kontrollierten Vokabularen, Thesauri und Normdaten als Linked Data sowie deren Anbindung an eigene Forschungsdaten einen weiteren bedeutenden Eckpunkt der Datenaufbereitung.

Die Gewährleistung einer erfolgreichen Strukturierung und Bereitstellung von Forschungsdaten innerhalb der Geisteswissenschaft im Sinne von Web 3.0 hängt im Wesentlichen von der Verfügbarkeit und Akzeptanz sogenannter **Virtueller Forschungsumgebungen** und **digitaler Forschungsinfrastrukturen**, die den Wissenschaftlern einen leichten, intuitiven und Mehrwert versprechenden Zugang zum eigenen Forschungsthema im Kontext von Linked Data anbieten.

Vor dem Hintergrund des diesjährigen DHd-Tagungsthemas der **Nachhaltigkeit** möchten wir uns der CIDOC CRM referenzierten Datenmodellierung und den Virtuellen Forschungsumgebungen sowie ihren Modulen und Features widmen. Das Panel nimmt sich vier laufende Forschungsprojekte auf dem Gebiet der **objekt- und raumbezogenen Forschung** zum Anlass aus praktischer Erfahrung multiperspektivisch zu berichten. Dabei wollen wir die E-CRM Entwickler aus DFG-geförderten WissKI I und II Projekt (2009-12, 2014-16) und die Anwender aus sich in unterschiedlichen Stadien befindenden geisteswissenschaftlichen Forschungsprojekten zur Sprache kommen lassen. Zum Ausdruck soll u. a. die Herausforderung der nicht konvergierenden Zielsetzung einzelner Forschungsprojekte kommen, deren Forschungsdaten jedoch im Sinne von Linked Data in der Praxis zusammengeführt werden sollen. Darüber hinaus wollen wir die Schwierigkeiten bei der Entwicklung von einzelnen Features und Modulen unter der Zielsetzung “einen gemeinsamen Weg zu gehen” offenlegen und mögliche Vorgehensweisen für die Zukunft projizieren. Anschließend wollen die die Zukunftsfähigkeit von WissKI-basierten (OWL DL / Graphdatenbank) Forschungsumgebungen und anderen Ansätzen (MySQL/Relationale Datenbank) in der Diskussionsrunde besprechen.

Virtuelle Rekonstruktionen in transnationalen Forschungsumgebungen – Das Portal:

Schlösser und Parkanlagen im ehemaligen Ostpreußen (ViReBa), 2013-2016

Piotr Kuroczyński

Das ViReBa-Projekt untersucht den gesamten Prozess der 3D-Computer-Rekonstruktion verloren gegangener Kunst und Architektur. Die vorläufigen Ergebnisse basieren auf der digitalen 3D-Rekonstruktion zerstörter ostpreußischer Barockschlösser (Schlodien, Friedrichstein) und bringen neue Erkenntnisse für die Quellenerschließung, Dokumentation, semantische Modellierung und Visualisierung von 3D-Datensätzen innerhalb der WebGL-Technologie. Der Schwerpunkt liegt dabei auf der Entwicklung eines menschen- und maschinenlesbaren Datenmodels zur Annotation und Integration diverser Meta- und Paradaten einschließlich der semantischen Auszeichnung von 2D und 3D-Datensätzen. Für kollaborative, interdisziplinäre und internationale Forschung bei und an der digitalen 3D-Rekonstruktion wird das CIDOC-CRM-basierte Framework von WissKI als Virtuelle Forschungsumgebung (VFU) seit 2014 adaptiert. Der Impulsvortrag zeigt kritisch die Erfahrungen, Herausforderungen und Potenziale, bei der Einrichtung einer VFU für die digitale hypothetische 3D-Rekonstruktion, die Dokumentation und Archivierung der Forschungsdaten und ihrer Derivate (u. a. im “Virtuellen Museum”).

Forschungsinfrastruktur Kunstdenkmäler in Ostmitteleuropa (FoKO), 2014-2017

Ksenia Stanicka-Brzezicka

Digitale Datenbanken sind in den letzten Jahren in der Erschließung von Kunstobjekten aller Art zum Standard geworden. Dabei ändern sich die technischen Möglichkeiten schnell und die Anpassung der Praktiken und Methoden der Kunstgeschichte stellt erhebliche Herausforderungen. Trotzdem entstehen viele neue kunsthistorische Datenbanken, vor allem in Rahmen von kurzfristigen Projekten, deren Nachhaltigkeit nicht garantiert ist. „Die Kunstgeschichte als Disziplin [hat] es bisher verpasst, neue methodische Grundlagen im Sinne einer nachhaltigen digitalen Quellenkritik bereitzustellen“ – lautet das Urteil in der Einleitung vom Summer Institut “Digital Collections” 2016 in Zürich/Lausanne (http://digital-collections.online/).

Das FoKO-Projekt, ein internationales Verbundprojekt, das den Aufbau einer interaktiven kunsthistorischen Forschungsinfrastruktur zum Ziel hat, stellt jedoch die Frage der Nachhaltigkeit stark in den Fokus. Im Austausch mit weiteren WissKI-Projekten (ViReBa, CbDD) strebt es nach der Entwicklung eines Datenmodells, das prototypenhaft für Foto- und Kunstdatenbanken verschiedene Entitäten, wie Kunstobjekte und Fotografien, zum einen einzeln, zum anderen Ihrer Eigenschaft der technischen Vervielfältigung nach multipel erfassen und beschreiben kann. Den Schwerpunkt des Projektes stellt die Entwicklung eines Datenmodells, das insgesamt nutzbar und übertragbar sein kann.

Corpus der barocken Deckenmalerei in Deutschland (CbDD), 2015-2040

Werner Köhler

Das Akademie-Projekt ist Mitte 2015 gestartet und hat die umfassende kunsthistorische Erforschung, Dokumentation und Präsentation der zwischen 1550 und 1800 entstandenen Werke der Wand- und Deckenmalerei auf dem Gebiet der Bundesrepublik Deutschland zur Aufgabe, wobei bis zum Jahr 2040 mehr als 5.000 bekannte Objekte dokumentiert werden.

Die lange Projektdauer ermöglicht die prototypische Entwicklung einer virtuellen Forschungsumgebung (VFU) für die Domäne der Kunstgeschichte insgesamt. Die Sicherstellung der Nachhaltigkeit einer solchen Entwicklung stellt eine zentrale Aufgabe der IT-Planung und des IT-Projektmanagements dar.

Aktuell wird das CIDOC-CRM-basierte VFU-Framework WissKI eingesetzt und vor dem Hintergrund der ISO-Qualitätsmodelle zum Software Engineering (ISO/IEC 9126, ISO/IEC 25000) hinsichtlich Funktionalität, Zuverlässigkeit, Benutzbarkeit, Effizienz, Änderbarkeit und Übertragbarkeit der Software evaluiert.

Im Panel sollen die Projekterfahrungen mit WissKI seit Oktober 2015 konkret dargestellt und mit den Erfahrungen aus den anderen WissKI-Projekten verglichen und diskutiert werden. Darüber hinaus sollen die Rahmenbedingungen für die kontinuierliche Anpassung, Erweiterung und Weiterentwicklung eines grundlegenden VFU-Frameworks für die Digital Humanities und die Entwicklung und Verstetigung einer nachhaltigen Infrastruktur thematisiert sowie Wege zu deren Realisierung aufgezeigt werden.

Topographie in Raum und Zeit: Ein digitales Raum-Zeit-Modell für vernetzte Forschung am Beispiel Nürnberg (TOPORAZ), 2015-2018

Armand Brahaj

TOPORAZ focusses on the topography of a quarter of historical Nuremberg, which is displayed at three to four time levels: the early modern period; (1870), 1939; and the present. The representation consists of geo-referenced 2D maps and 3D models and a factual database covering buildings, furnishing, iconography,

persons, and social networks. Maps and 3D models serve as a structure for navigation and help visualizing results from database queries. The information is maintained in a relational database which implements a semantic data model heavily influenced by CIDOC CRM. This approach enables researcher to query for facts like 'Who inhabited a building at a given time?', 'How did a building evolve through history?' or 'Who donated this statue and where is it located today?'

TOPORAZ directly links 3D objects of the interactive city model (e.g. streets, buildings, floors and rooms) to research literature and source material (texts, images, sound) via hotspots. The Virtual Research Environment (VRE) presents those materials to users based on their virtual location within the model and the chosen time level. The VRE supports interdisciplinary research approaches and transdisciplinary networking, brings together researchers from art history, architecture, 3D modelling and computer science.

WissKI im Museum – Einsatzszenarien im Germanischen Nationalmuseum

Mark Fichtner

Das Germanische Nationalmuseum (GNM) vereint als größtes kulturgeschichtliches Museum des deutschen Sprachraums vielfältige Sammlungen und Archive, das Institut für Kunsttechnologie und Konservierung sowie die größte öffentlich zugängliche Spezialbibliothek für deutsche Kulturgeschichte. Die Forschungseinheiten führten aus historischen Gründen und bedingt durch verschiedene Erschließungskonventionen zu spartenspezifischen, an die Anforderungen angepassten Datenbanksystemen. Daraus resultieren nachhaltige Probleme, so sind die Daten trotz ähnlicher Nutzerkreise und sich überschneidender, ergänzender Inhalte nur schwer austauschbar, kaum verknüpfbar und nicht homogen durchsuchbar.

Zur Lösung dieses Problems wurde im DFG geförderten Projekt „Wissenschaftliche KommunikationsInfrastruktur" (WissKI) eine Software entwickelt, die eine ideale Plattform für Linked Open Data bietet. Auf Basis von ISO 21127 (CIDOC CRM) als gemeinsame Lingua Franca bleibt die Interpretierbarkeit der Inhalte gewährleistet, während das System durch Domänenontologien an die jeweiligen Fachbereiche angepasst werden kann.

Der Vortrag stellt WissKI, das seit 2013 am GNM im stetigen Einsatz in nahezu allen Forschungs- und Ausstellungsprojekten ist, aus der Sicht der Informatik vor. Das häufigste der drei Nutzungsszenarien ist der Einsatz als virtuelle Forschungsinfrastruktur, die Kernaufgabe für die das System auch konzipiert wurde. Weiterhin dient es als Softwareplattform für virtuelle Ausstellungen und als einheitliches Ausstellungs- und Planungstool.

Anforderungen an nachhaltige Entwicklung von Software für Forschungsinfrastrukturen

Barbara Fichtl

Aufbauend auf den Erfahrungen der im Panel vorgestellten Projekte stellt der abschließende Beitrag die Frage nach der Nachhaltigkeit von Software-Entwicklung im Bereich der Digital Humanities. Welche Rahmenbedingungen sind nötig, um Forschungsinfrastrukturen langfristig zu betreiben? Was sollte bei der Projektentwicklung und -durchführung hinsichtlich der Nachhaltigkeit beachtet werden? Wie müsste eine Projektförderung aussehen, die nachhaltige Software-Entwicklung und den langfristigen Betrieb von Forschungsinfrastrukturen unterstützt?

Bibliographie

Berners-Lee, Tim / Hendler, James / Lassila, Ora (2001): „The Semantic Web", in: *Scientific American* 34–43.

Caraffa, Constanza (2011): „‚Wenden!' Fotografien in Archiven im Zeitalter ihrer Digitalisierbarkeit: ein ‚materialturn'", in: *Rundbrief Fotografie* 18, 3: 8–15.

Cellary, Wojciech / Walczak, Krzysztof (2012): *Interactive 3D Multimedia Content: Models for Creation, Management, Search and Presentation*. Heidelberg: Springer.

Bentkowska-Kafel, Anna / Denard, Hugh / Baker, Drew (2012): *Paradata and Transparency in Virtual Heritage*. London: Ashgate.

Rat für Informationsinfrastrukturen (2016): *Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland* http://www.rfii.de/de/category/dokumente/ [letzter Zugriff 25. August 2016].

Münster, Sander / Pfarr-Harfst, Mieke / Kuroczyński, Piotr / Ioannides, Marinos (2016): *How to manage data and knowledge related to interpretative digital 3D reconstructions of Cultural Heritage?*. Heidelberg: Springer LNCS.

Kuroczyński, Piotr / Bell, Peter / Dieckmann, Lisa (2016): *Computing Art Reader: Einführung in die digitale Kunstgeschichte*. Arthistoricum.net-ART-Books, Heidelberg (in Edition).

Zugänglichkeit und dauerhafte Nutzbarkeit historischer Bildrepositorien für Forschung und Vermittlung

Niebling, Florian
florian.niebling@uni-wuerzburg.de
Julius-Maximilians Universität Würzburg, Deutschland

Münster, Sander
sander.muenster@tu-dresden.de
Technische Universität Dresden, Deutschland

Friedrichs, Kristina
kristina.friedrichs@uni-wuerzburg.de
Julius-Maximilians Universität Würzburg, Deutschland

Henze, Frank
frank.henze@b-tu.de
Brandenburgische Technische Universität Cottbus-Senftenberg, Deutschland

Kröber, Cindy
cindy.kroeber@tu-dresden.de
Technische Universität Dresden, Deutschland

Bruschke, Jonas
jonas.bruschke@uni-wuerzburg.de
Julius-Maximilians Universität Würzburg, Deutschland

Digitalisate historischer Fotografien und deren Nutzbarkeit zur geschichtswissenschaftlichen Forschung und quellenbasierten Vermittlung stellen ebenso wie räumliche Modelle historischer Objekte Kernthemen der Digital Humanities dar. Angesichts des Umfangs derartiger Repositorien besteht eine wesentliche Herausforderung darin, für die Beantwortung geschichtswissenschaftlicher Fragestellungen relevante und aussagekräftige Quellen zu finden, zu kontextualisieren sowie die darin beschriebenen historischen Objekte vorstellbar zu machen. Die Verbindung zwischen digitalen Bildrepositorien und Raumbezug verspricht durch eine Zusammenführung und nutzerzentrierte Präsentation von Informationsbeständen ein umfassendes Repertoire technischer Unterstützungsoptionen geschichtswissenschaftlicher Forschungspraxis. Im Gegensatz zu bisherigen Zugängen zu Bild- und Planrepositorien wird durch die dreidimensional-räumliche Verortung von Quellen, ebenso wie durch ihre Vor-Ort-Präsentation ein hohes Maß intuitiver Zugänglichkeit und Kontextbezuges geschaffen. Im Panel diskutiert werden innovative Softwarewerkzeuge und damit verbundene methodische Ansätze für die Verwendung historischer Bildrepositorien in der stadt- und architekturgeschichtlichen Forschung. Hierbei sollen ausgehend von aktuellen digitalen Rekonstruktionsprojekten Forschungsmethoden vorgestellt, kategorisiert und hinsichtlich vorhandener Unterstützungsbedarfe diskutiert werden. Davon ausgehend werden softwaretechnische Methoden aufgezeigt, welche einerseits den Zugang zu Bildrepositorien erleichtern und dadurch eine dauerhafte Benutzbarkeit sicherstellen sollen, sowie andererseits in Fotografien verborgenes Wissen, beispielsweise über den Betrachterstandpunkt und den Zeitpunkt der Aufnahme zugänglich machen.

Dr. Kristina Friedrichs: Methoden architekturgeschichtlicher Forschung

Die Kunstgeschichte kann auf eine lange Tradition der wissenschaftlichen Auseinandersetzung mit Architektur zurückblicken. Im Zuge dessen haben sich verschiedene Methoden des Herangehens entwickelt, die sowohl tatsächlich erhaltene als auch nie gebaute oder später zerstörte Bauwerke zum Zwecke der Chronologisierung, der historischen Kontextualisierung und Bedeutungsentschlüsselung erschließen.

Neue technologische Möglichkeiten erlauben es Architekturhistorikern einerseits, ihre Untersuchungen auf einen größeren Fundus an Quellen aufzubauen, die beispielsweise durch digitale Bildarchive zur Verfügung gestellt werden. Andererseits ergeben sich neue methodische Ansätze aus innovativen Software-Werkzeugen, die helfen, die Quellen zeitlich wie räumlich zu verorten, oder die Forschung durch Visualisierungen bei der Erstellung von Datierungen, stilkritischen Betrachtungen, der Zuweisung von Autorenschaften oder bauarchäologischen Untersuchungen zu unterstützen (Verstegen 2007).

Gerade am Beispiel der Stadt Dresden mit ihrer reichen und wechselhaften Geschichte lassen sich dank umfangreicher Bildrepositorien neue Untersuchungsfelder eröffnen. Am Dresdner Zwinger wurden große Teile der Planungs- und Baugeschichte durch Visualisierungen nachvollzogen und darüber hinaus die fertigen Modelle in die Vermittlung innerhalb eines musealen Kontextes überführt (Jahn/Welich 2009). Für die Kunstgeschichte ergeben sich vor diesem Hintergrund mannigfaltige neue Arbeitsansätze, die sowohl hinsichtlich ihrer Methodik diskutiert werden müssen, als auch einer Unterstützung mithilfe von adäquaten Werkzeugen aus den technischen Disziplinen bedürfen.

Dr. Sander Münster: Eine Wissensbasis für die Digital Visual Humanities

Eine daran eng anknüpfende Frage ist die nach einer methodischen Validierung digitaler Methoden sowie insbesondere der Verwendung von Bildrepositorien im Kontext der Architekturgeschichte (c.f. Arbeitstagung digitale Kunstgeschichte 2014). Dies umfasst zunächst einmal den Bedarf, ein Spektrum digitaler Werkzeuge sowie Verwendungskontexte im Kontext der Kunstgeschichte zu systematisieren (Kohle 2013, Heusinger 1989). Vor diesem Hintergrund sollen im Rahmen dieses Vortrags Ergebnisse dreier Workshops vorgestellt werden, welche 2016 auf internationalen Konferenzen abgehalten wurden und bei welchen unter Einbeziehung von ca. 100 Experten mit den Schwerpunkten Cultural Heritage und Digital Visual Humanities wesentliche Methoden und Forschungsansätze sowie Podien erfasst und systematisiert wurden.

Darüber hinaus sollen im Vortrag exemplarisch spezifische fachkulturelle sowie wissenschaftlich-methodische Herausforderungen des Einsatzes digitaler Methoden sowie insbesondere von Bildrepositorien im Kontext architekturgeschichtlicher Forschung beleuchtet werden. Dazu gehören Aspekte wie die Transparentmachung von Erkenntnisprozessen (Benkowska-Kafel et al. 2012) ebenso wie eine bildgestützte Diskurskultur (vgl. Münster, Friedrichs & Hegel in Vorb.) sowie nicht zuletzt der Blick auf eine digitale Nachhaltigkeit. Im Ergebnis sollen somit nicht nur ein methodologischer State-Of-the-Art vorgestellt, sondern auch die Determinanten für die Konzeption digitaler Werkzeuge und Unterstützungsoptionen skizziert werden

Cindy Kröber: Zielgruppen-orientierte Erstellung von Werkzeugen für die Arbeit mit Bildrepositorien

Der Erfolg von Bilddatenbanken hängt stark von der Usability der Anwendung sowie der Tauglichkeit als Forschungs- oder Vermittlungstool ab. Bisherige Werkzeuge und Funktionalitäten entsprechender Anwendungen entsprechen häufig nicht den Bedarfen der architektur- und kunstgeschichtlichen Forschung und Vermittlung (Dudek et al. 2015).

Allgemeine Anforderungen der Nutzer sind ein schnelles Verstehen der Daten und Informationen, effiziente Such- und Filterfunktionen und eine intuitiv bedienbare Softwareoberfläche und Navigation (Barreau et al. 2014). Für Forschungsanliegen spielen wissenschaftliche Standards wie die ausführliche Dokumentation durch Metadaten eine wichtige Rolle (Maina/Suleman 2015). Eine interessierte Öffentlichkeit erwartet hingegen eine direkte und überschaubar gestaltete Einführung in das Thema und die entsprechenden Daten (Maina/Suleman 2015) sowie weitere Informationsangebote nach Bedarf. Für die Forschung sind visuelle Darstellungen von Hypothesen und Zusammenhängen wichtig (López-Romero 2014). Die erweiterte Bildanalyse von Fotos eines Objektes über die Zeit erlaubt die Detektion baulicher Veränderungen.

Um zielgruppen-orientiert Softwarewerkzeuge für die Arbeit mit Bildrepositorien und insbesondere Bilddatendanken zu entwickeln, müssen die Unterstützungsmöglichkeiten identifiziert, konzipiert und überprüft werden. Die Nutzer sind von Beginn an mit Hilfe qualitativer Interviews und umfassenden Untersuchungen zu Nutzerverhalten und Nutzerinteraktion involviert.

Jonas Bruschke: Werkzeuge für die Dokumentation digitaler Rekonstruktionsprozesse

Digitale Rekonstruktionen können Experten und Laien ein Bild nicht mehr oder nur in Teilen existenter Gegenstandände vermitteln. 3D-Modelle sind dabei nicht nur Gegenstand der Betrachtung, sondern auch Forschungsgegenstände. Neben den materiellen Quellen die bei der Erstellung von 3D-Modellen eingesetzt werden, wie Pläne und Fotografien, handelt es sich oft auch um immaterielle Quellen, beispielsweise die Entscheidung von Experten. Resultierende Visualisierungen haben letztendlich aber keinen direkten Bezug mehr zu den verwendeten Quellen. In aller Regel ist für eine externe, nicht an der Entstehung des Modells beteiligte Person oft nur schwer nachvollziehbar, ob eine Rekonstruktion auf verlässlichen Fakten

beruht und inwieweit und welche Hypothesen bei der Erstellung eine Rolle spielten. Eine ausführliche, lückenlose Dokumentation der Rekonstruktion ist daher essentiell und sollte möglichst alle Aspekte und jegliches während der Bearbeitung erlangte Wissen umfassen. Dies betrifft nebst der Protokollierung der Entscheidungen auch Schwierigkeiten während des Entstehungsprozesses.

Eine solch umfangreiche Dokumentation kommt in den Rekonstruktionsprojekten in der Regel nicht zustande (Pfarr 2010, Münster 2014). Zur Unterstützung des Dokumentationsverhaltens müssen interdisziplinären Projektteams, vorrangig bestehend aus Historikern und Modelleuren, geeignete Werkzeuge in die Hand gelegt werden. Die Abläufe und Problemstellungen solcher Projekte wurden bereits umfangreich untersucht (Münster 2014). Darauf aufbauend wurde ein erster Prototyp entwickelt (Bruschke 2015), welcher zum einen als zentrales Element während eines Projektes zum Einsatz kommen soll, indem es von der Koordination des Projektes über das Einpflegen und Halten der Daten bis hin zur direkten Arbeit und Diskussion am 3D-Modell viele Abläufe eines Rekonstruktionsprojektes unterstützt und gleichzeitig auch protokolliert. Dieses angesammelte Wissen kann außenstehenden Personen in Form einer Rechercheplattform zugänglich gemacht werden und gegebenenfalls durch sie verifiziert werden.

Dr. Frank Henze: Photogrammetrische Methoden zur Wissensgenerierung aus Bildbeständen

Das Potenzial fotografischer und photogrammetrischer Aufnahmen reicht von der reinen *Bilddokumentation* im Bereich der Archäologie und Denkmalpflege, über die *Bildinterpretation*, zum Beispiel für Schadensdokumentationen, bis hin zur Erstellung *maßstäblicher Bildpläne* und komplexer *3D-Modelle* für baugeschichtlich-archäologische Untersuchungen (z.B. Bührer et al. 2001, Hanke 2001).

Aus fotografischen Aufnahmen lassen sich, bei Vorliegen entsprechender Bildinhalte, geometrische Informationen über die abgebildeten Objekte zum Zeitpunkt der Aufnahme rekonstruieren. Die Grundlagen für die geometrische Rekonstruktion aus historischen Fotografien bilden die analytischen Verfahren der Photogrammetrie, d.h. die Gewinnung zwei- und dreidimensionaler Objektgeometrien aus den zweidimensionalen Bildinformationen. Beispiele für die photogrammetrische Auswertung historischer Aufnahmen und Messbilder finden sich unter anderem in Wiedemann et al. 2000, Bräuer-Burchardt und Voss 2001, Henze et al. 2009 oder Siedler et al. 2011. Die klassischen Verfahren der analytischen Photogrammetrie werden dabei zunehmend ergänzt durch angepasste Verfahren der digitalen Bildverarbeitung und Bildanalyse. Der aufwändige Prozess der manuellen Bildauswertung kann damit weitgehend automatisiert werden, womit auch große Bildbestände für eine automatische Gewinnung geometrischer Informationen erschlossen werden können (Pomaska 2011).

Bisher werden automatisierte photogrammetrische Verfahren in der Regel jedoch ausschließlich für die Auswertung aktueller, zumeist digitaler Aufnahmen eingesetzt. Angepasste Verfahren für eine (semi-) automatische Auswertung historischer Bildbestände fehlen bisher. Dabei muss u.a. auf die Besonderheiten gescannter Analogaufnahmen mit zumeist unbekannter Kamerageometrie, fehlenden bzw. minimalen Objektinformationen und z.T. geringer radiometrischer und geometrischer Auflösung reagiert werden. Ziel ist es, anwendungsorientierte Werkzeuge für eine photogrammetrische Auswertung historischer Fotografien zu entwickeln und diese in den Prozess der geschichtswissenschaftlichen Bildanalyse zu integrieren und damit einen räumlichen Bezug zur heutigen Situation zu schaffen.

Dr.-Ing. Florian Niebling: Augmented Reality in den Visual Humanities

Bei der Nutzung digitaler Bildrepositorien sind zwei wesentliche Vorgehensweisen der Informationserschließung erkennbar: Einerseits ein selbstgesteuertes Durchsuchen von Sammlungen historischer Fotografien, Zeichnungen und Pläne, andererseits eine orts- oder kontextbezogene Informationsvermittlung beispielsweise im Zuge stadträumlicher oder musealer Präsentation (Münster et al. 2016). Die Vor-Ort-Darstellung von und Interaktion mit geschichtswissenschaftlichen Daten in der Augmented Reality hat hierbei in den letzten Jahren an Bedeutung gewonnen und wurde vielfältig erprobt und untersucht (Livingston et al. 2008; Zöllner et al. 2010; Walczak 2011).

Augmented Reality beschreibt dabei die Anreicherung der realen Welt durch virtuelle Daten, wobei es sich sowohl um 3D-Modelle, Texte, Bilder, Filme oder auch Audiodaten handeln kann. Durch die Anreicherung der Realität oder Ersetzung von Teilen der Realität können Augmented Reality Methoden helfen den Unterschied zwischen verschiedenen

Zuständen von Objekten darzustellen (Niebling, 2008). Im geschichtswissenschaftlichen und stadthistorischen Kontext wird es dem Betrachter ermöglicht, interaktiv visuelle und textuelle Informationen zu dreidimensional vermessenen Objekten in ihrem historischen räumlichen Bezugssystem zu erfassen.
Ein Hauptaugenmerk liegt dabei auf der Zugänglichkeit historischer Datenbestände. Wie können Interaktionsmöglichkeiten mit virtuellen Gebäuden und mit ihnen verknüpften Informationen gestaltet werden? Können aus dem Umgang mit Mobilgeräten bekannte Interaktionsmetaphern in der Augmented Reality weiterverwendet werden? Welche Vermittlungsmethoden können in Augmented Reality Anwendungen zum Einsatz kommen?

Bibliographie

Barreau Jean-Baptiste / Gaugne, Ronan / Bernard, Yann / Le Cloirec, Gaétan / Gouranton, Valérie (2014): „Virtual reality tools for the West Digital Conservatory of Archaeological Heritage“, in: *Proceedings of the 2014 Virtual Reality International Conference* 1–4.

Bentkowska-Kafel, Anna / Denard, Hugh / Baker, Drew (2012): *Paradata and Transparency in Virtual Heritage*. Burlington: Ashgate.

Bräuer-Burchardt, Christian / Voss, Klaus (2001): „Facade Reconstruction of Destroyed Buildings Using Historical Photographs“, in: Albertz, Jörg (ed.): *Proceedings of the XVIII. International CIPA Symposium* 543–550.

Bruschke, Jonas (2015): *DokuVis – Ein Dokumentationssystem für Digitale Rekonstruktionen*. Masterarbeit, HTW Dresden.

Bührer, Thomas / Grün, Armin / Zhang, Li / Fraser, Clive / Rüther, Heinz (2001): „Photogrammetric Reconstruction and 3D Visualization of Bet Gorgis, a Rock-hewn Church in Ethiopia“, in: Albertz, Jörg (ed.): *Proceedings of the XVIII. International CIPA Symposium* 338–344.

Dudek, Iwona / Blaise, Jean-Yves / De Luca, Livio / Bergerot, Laurent / Renaudin, Noémie (2015): „How was this done? An attempt at formalising and memorising a digital asset's making-of“, in: *Digital Heritage* 2: 343–346.

Gabbard, Joseph L. / Swan, J. Edward (2008): „Usability Engineering for Augmented Reality: Employing User-Based Studies to Inform Design“, in: *IEEE Transactions on Visualization and Computer Graphics*, 14 (3): 513–525.

Henze, Frank / Lehmann, Heike / Bruschke, Bettina (2009): „Nutzung historischer Pläne und Bilder für die Stadtforschungen in Baalbek / Libanon“, in: *Photogrammetrie - Fernerkundung - Geoinformation* 3/2009: 221–234.

Hertzig, Stefan / Friedrichs, Kristina (i. Vorb.): *Das Japanische Palais in Dresden: Vom Porzellanschloss Augusts des Starken zu einem Museum des Bildungsbürgertums*.

Heusinger, Lutz (1989): „Applications of Computers in the History of Art“, in: Hamber, Anthony / Miles, Jean / Vaughan, William (eds.): *Computers and the History of Art*. London: Mansell Pub 1–22.

Internationale Arbeitstagung „Digitale Kunstgeschichte: Herausforderungen und Perspektiven“ (2014): *Zürcher Erklärung zur digitalen Kunstgeschichte*.

Jahn, Peter Heinrich / Welich, Dirk (2009): „Zurück in die Zukunft: die Visualisierung planungs- und baugeschichtlicher Aspekte des Dresdner Zwingers“, in: *Jahrbuch Staatliche Schlösser, Burgen und Gärten Sachsen* 16: 51–72.

Kohle, Hubertus (2013): *Digitale Bildwissenschaft*. Glückstadt.

Livingston, Mark A. / Bimber, Oliver / Saito, Hideo (2008): *Proceedings of the 7th IEEE International Symposium on Mixed and Augmented Reality*. Cambridge, UK. / Piscataway, N.J.: IEEE Xplore.

López-Romero, Elías (2014): „Out of the box: exploring the 3D modelling potential of ancient image archives“, in: *Virtual archaeology review* 5 (10): 107–116.

Maina, Job King'ori / Suleman, Hussein (2015): „Enhancing Digital Heritage Archives Using Gamified Annotations“, in: *Digital Libraries: Providing Quality Information* 9469. Seoul: 169-179.

Münster, Sander (2014): *Interdisziplinäre Kooperation bei der Erstellung virtueller geschichtswissenschaftlicher 3D-Rekonstruktionen*. Dissertation, TU Dresden.

Münster, Sander / Niebling, Florian (2016): „HistStadt4D - Multimodale Zugänge zu historischen Bildrepositorien zur Unterstützung stadt- und baugeschichtlicher Forschung und Vermittlung“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 203-208.

Münster, Sander / Friedrichs, Kristina / Hegel, Wolfgang (eingereicht): „3D Reconstruction techniques as a Cultural Shift in Art History?“, in: *International Journal of Digital Art History*.

Niebling, Florian / Griesser, Rita T. / Woessner, Uwe (2008): „Using Augmented Reality and Interactive Simulations to Realize Hybrid Prototypes“, in: *Advances in Visual Computing, 4th International Symposium* ISVC 2008. Proceedings I: 1008–1017.

Pfarr, Mieke (2010): *Dokumentationssystem für Digitale Rekonstruktionen am Beispiel der Grabanlage Zhaoling, Provinz Saanxi, China*. Dissertation, TU Darmstadt.

Pomaska, Günter (2011): „Zur Dokumentation und 3D-Modellierung von Denkmalen mit digitalen fotografischen Verfahren“, in: Heine, Katja / Rheidt, Klaus / Henze, Frank / Riedel, Alexandra (eds.): *Von Handaufmaß bis High Tech III – 3D in der historischen Bauforschung*. Mainz: Verlag Philipp von Zabern 79–84.

Siedler, Gunnar / Sacher, Gisbert / Vetter, Sebastian (2011): „Photogrammetrische Auswertung historischer Fotografien am Potsdamer Stadtschloss“, in: Heine, Katja / Rheidt, Klaus / Henze, Frank / Riedel, Alexandra (eds.): *Von Handaufmaß bis High Tech III - 3D in der historischen Bauforschung*. Mainz: Verlag Philipp von Zabern 26–32.

Verstegen, Ute (2007): „Vom Mehrwert digitaler Simulationen dreidimensionaler Bauten und Objekte in der architekturgeschichtlichen Forschung und Lehre“, Vortrag am 16.3.2007, in: *XXIX. Deutscher Kunsthistorikertag*, Regensburg.

Walczak, Krzysztof / Cellary, Wojciech / Prinke, Andrzej (2011): „Interactive Presentation of Archaeological Objects Using Virtual and Augmented Reality“, in: Jerem, Erszébet / Redő, Ferenc / Szeverényi, Vajk (eds.): *On the Road to Reconstructing the Past*. Proceedings of the 36th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA). Budapest: Archaeolingua.

Wiedemann, Albert / Hemmleb, Matthias / Albertz, Jörg (2000): „Reconstruction of historical buildings based on images from the Meydenbauer archives“, in: *International Archives of Photogrammetry and Remote Sensing* XXXIII (B5/2): 887–893.

Zöllner, Michael / Becker, Mario / Keil, Jens (2010): „Snapshot Augmented Reality - Augmented Photography“, in: Artusi, Alessandro / Joly-Parvex, Morwena / Lucet, Genevieve / Ribes, Alejandro / Pitzalis, Denis (eds.): *11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2010).* Paris: Eurographics Association.

Vorträge

Ambige idiomatische Ausdrücke in kinderliterarischen Texten: Mehrwert einer Datenbankanalyse

Wagner, Wiltrud
wiltrud.wagner@uni-tuebingen.de
Eberhard Karls Universität Tübingen, Deutschland

In meinem Vortrag setze ich mich mit der Frage auseinander, welchen Beitrag die Datenbank TInCAP („Tübingen Interdisciplinary Corpus of Ambiguity Phenomena“), die bei der Tagung der Digital Humanities im deutschsprachigen Raum 2016 in Leipzig vorgestellt wurde und die der Sammlung und Annotation von Ambiguitätsbelegen dient, zur Erforschung des Phänomens „Ambiguität“ leisten kann. Den Mehrwert, den TInCAP durch die innovative interdisziplinäre Annotation und die Zusammenführung von Belegen in einer durchsuchbaren Datenbank liefert, werde ich am Beispiel ambiger idiomatischer Ausdrücke in kinderliterarischen Texten illustrieren.

Die Datenbank TInCAP entsteht im Rahmen des interdisziplinären Graduiertenkollegs *GRK 1808 Ambiguität – Produktion und Rezeption* (www.ambiguitaet.uni-tuebingen.de ; Arbeitsgruppe TInCAP: Wiltrud Wagner, Lisa Ebert, Jutta Hartmann, Gesa Schole, Susanne Winkler) mit dem Zweck, Ambiguitätsbelege aus allen beteiligten Disziplinen zu sammeln und zu annotieren. Hauptziele sind dabei die interdisziplinäre Auseinandersetzung mit dem Phänomen Ambiguität durch die Erstellung eines gemeinsamen Annotationsschemas sowie die nachhaltige Speicherung und Zugänglichmachung der Datensammlung für die nationale und internationale Forschungsgemeinschaft (in Kürze über die Homepage des GRK 1808).

Auch wenn alle an diesem Projekt beteiligten WissenschaftlerInnen das Interesse am Phänomen der Ambiguität verbindet, das hier als Doppel- oder Mehrdeutigkeit in ihren verschiedensten Formen verstanden wird, so sind die zu annotierenden Belege doch sehr divers: Durch die Vielzahl der beteiligten Disziplinen unterscheiden sich die Belege hinsichtlich Medium (aktuell: Schrift, Audio, Bild, Video) und Sprache (aktuell: Deutsch, Englisch, Französisch, Hebräisch, Italienisch, Latein, Spanisch, Griechisch), aber auch Umfang. Im Bestreben, eine gemeinsame Datenbank aufzubauen, sahen wir uns demnach zwei großen Herausforderungen gegenüber gestellt: (1) Der Erarbeitung einer disziplinenübergreifenden Terminologie, die einerseits präzise, andererseits aber nicht an das Vokabular einer der Disziplinen gebunden ist, und (2) der Entwicklung eines interdisziplinären Annotationsschemas, das – trotz der notwendigen Komplexitätsreduktion – den Anforderungen der einzelnen Disziplinen genügt und für alle Beteiligten profitabel ist.

Das Ergebnis ist ein Annotationsschema, das die folgenden fünf Punkte fokussiert:

Communication level: Auf welcher Ebene der Kommunikation wird die Ambiguität annotiert? Für literarische Texte wird zum Beispiel zwischen der Ebene der fiktiven Charaktere, der Ebene des/der Erzähler(s) und der Ebene des Autors und Lesers unterschieden.

Strategic or non-strategic production and/or perception: Wird die Ambiguität strategisch produziert? Wird die Ambiguität strategisch rezipiert?

Level of Trigger and Range: Zu annotieren ist, auf welcher Ebene die Ambiguität ausgelöst wird und bis zu welcher Ebene sie relevant ist. Die Ebenen für Auslöser und Wirkung der Ambiguität bilden dabei ein Größenverhältnis ab, analog zum menschlichen Körper, bei dem sich größere Elemente aus kleineren zusammensetzen (z.B. die Ebene *Subelement*, die u.a. Phoneme, Grapheme und Morpheme umfasst; die Ebene *Element*, die u.a. Worte umfasst; usw.).

Type of Paraphrase Relation: In welchem Verhältnis stehen die möglichen Lesarten zueinander? Sind sie voneinander abgeleitet oder völlig unabhängig voneinander?

Phenomenon: Welches Phänomen steht mit der vorliegenden Ambiguität im Zusammenhang? Hier kann und soll disziplininternes Vokabular zur Anwendung kommen, um die Einbindung in den jeweiligen Forschungskontext zu gewährleisten.

Zusätzlich ist die Verknüpfung von Annotationen möglich, zum Beispiel, wenn ein Beleg auf verschiedenen Kommunikationsebenen (unterschiedlich) annotiert wird.

Die Nachhaltigkeit der gesammelten Daten wird durch eine Kombination verschiedener Faktoren gewährleistet: Das von uns entwickelte XML-Schema ist soweit möglich TEI-konform, es wurde für die inhaltliche Annotation der Daten um ein eigenes Schema erweitert. Der gesamte Korpus bzw. Subkorpora können im XML-Format im- und exportiert werden. Diese XML-Dateien werden in Kooperation mit Clarin-D Tübingen im Rahmen der universitären Infrastruktur langfristig gespeichert, katalogisiert und mit PIDs zugänglich gemacht. Teilkorpora können dabei ebenso exportiert werden wie das Gesamtkorpus. Bei Video-, Audio- und Bilddateien halten wir uns an die üblichen Standards für nachhaltige Datenformate (nicht-proprietäre Formate, Formate mit gutem Nachnutzungswert).

Nach der allgemeinen Vorstellung der Datenbank wende ich mich im zweiten Teil des Vortrags der Frage zu, was die Datenbank im Hinblick auf konkrete Fragestellungen leistet. Die von mir in die Datenbank eingebrachten Ambiguitätsbelege entstammen zum größten Teil meiner Dissertation, die einen interdisziplinären Beitrag zur Ambiguitätsforschung leistet: Der linguistische Teil der Arbeit untersucht, wie idiomatischen Ausdrücken das Potential zur Ambiguität inhärent sein kann. An der Schnittstelle zur Literaturwissenschaft zeigt die Arbeit, wann und wie idiomatische Ausdrücke in Interaktion mit unterschiedlichen Kotexten ihr Ambiguitätspotential entfalten. Am Beispiel von kinderliterarischen Texten wird schließlich dargestellt, wie die aus dieser Interaktion resultierende Bewusstmachung von Ambiguität als sprachspielerisches Potential für literarische Texte produktiv gemacht werden kann. (a)-(c) stellen typische Beispiele aus meinem Korpus dar, die jeweils annotierten Stellen sind fett markiert:

(a)

One day he went to King Big-Twytt, who was eating a bathtub of roast chicken, custard and chips, and said: 'King - I want a licence to catch ye dragons.'

'What?' said King Twytt. 'But ye dragons are dangerous! They eat ye farm animals.'

'So do we,' said Sir Nobonk, 'and no one says we're dangerous.'

'Yea, very well,' said King Twytt, 'I will give you a licence, but **be it on your own head**.'

So Sir Nobonk strapped the licence to his head.

Sir Nobonk had been in many wars. Usually […]

(Spike Milligan: *Sir Nobonk and the terrible, awful, dreadful, naughty, nasty Dragon*, 1982)

(b)

Draw the drapes *when the sun comes in*.

read Amelia Bedelia. She looked up. The sun was coming in. Amelia Bedelia looked at the list again. "Draw the drapes? That's what it says. I'm not much of a hand at drawing, but I'll try."

So Amelia Bedelia sat right down and she drew those drapes.

(Peggy Parish: *Amelia Bedelia*,1963.)

(c)

Tom ging auf den frierenden König zu. „Ich bin gekommen, um mein Versprechen einzulösen“, sagte er und warf die Satteltasche auf den Tisch.

König Knöterich schaute ungläubig auf die Tasche. „Hast du mir etwa ein Paar warme Handschuhe mitgebracht?“

„Nein, Herr König“, antwortete Tom. „Etwas viel Kostbareres. Ich habe für Euch den goldenen Dings, äh, Kelch erobert.“

„Aahhh! Oohhh!“, hallte es durch den Saal.

„Ihr wollt wohl den König **auf den Arm nehmen**“, sagte Friedrich von Edelstein.

„Ich fürchte, mit den vielen Umhängen und Mützen ist mir der König zu schwer“, grinste Tom.

(Bernd Schreiber: Ritter Tollkühn und der goldene Dings, 2010.)

Die Annotation meiner Beispiele mit TInCAP ermöglicht die Sichtbarmachung von Aspekten, die bei der reinen linguistischen oder literaturwissenschaftlichen Analyse möglicherweise verborgen bleiben. Besonderes Gewicht kommt dabei der Möglichkeit zu, Ambiguitäten auf mehreren Kommunikationsebenen zu annotieren und die resultierenden Annotationen zu verknüpfen. Dies möchte ich anhand von Beispielen wie (a)-(c) illustrieren und mich dabei auf folgende Phänomene konzentrieren:

strategische vs. nicht-strategische Produktion/ Rezeption: In den untersuchten kinderliterarischen Texten erfolgt meist die Produktion auf der innersten Ebene (Ebene der Figuren) nicht strategisch, auf der

äußersten Ebene (Ebene des Autors) jedoch strategisch.

Typ der Ambiguitätsverwendung: Sehr häufig wird in den untersuchten kinderliterarischen Texten die Ambiguität auf der innersten Ebene nicht erkannt, auf der äußersten Ebene muss jedoch eine semantische Reanalyse erfolgen, wodurch die Ambiguität sichtbar gemacht wird.

Erste Lesart (phrasal vs. kompositional): Die erste (und damit oftmals einzige) Lesart auf der innersten Ebene ist sehr häufig die kompositionale. Auf der äußersten Ebene ist es jedoch die phrasale Lesart, die primär verarbeitet wird, woraus die Notwendigkeit der semantischen Reanalyse resultiert.

Diese Phänomene, die erst durch die Annotation mit TInCAP und durch entsprechende Suchabfragen sichtbar werden, zeigen das Potential, das diese Datenbank innerhalb eines Projekts entfaltet. In einem abschließenden Ausblick möchte ich darüber hinaus auf den interdisziplinären Nutzen der Datenbank verweisen, der im Rahmen des GRK 1808 bereits zum Tragen kommt, insbesondere in der Vergleichbarkeit, die über Medien hinweg geschaffen wird.

Bibliographie

Hartmann, Jutta / Sauter, Corinna / Schole, Gesa / Wagner, Wiltrud / Gietz, Peter / Winkler, Susanne (2016): *TInCAP – ein interdisziplinäres Korpus zu Ambiguitätsphänomenen*. Posterpräsentation, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung*.

Hartmann, Jutta / Ebert, Lisa / Schole, Gesa / Wagner, Wiltrud / Winkler, Susanne (eingereicht): „Annotating Ambiguity Across Disciplines: The Tübingen Interdisciplinary Corpus of Ambiguity Phenomena“, in: Bauer, Matthias / Zirker, Angelika (eds.): *Strategies of Ambiguity*.

Hartmann, Jutta / Ebert, Lisa / Schole, Gesa / Wagner, Wiltrud / Winkler, Susanne (in Vorbereitung): *TInCAP User Manual*.

Klein, Wolfgang / Winkler, Susanne (eds.) (2010): *Ambiguität*. Zeitschrift für Literaturwissenschaft und Linguistik 40 (158). Stuttgart: Metzler.

TEI Consortium (eds.): *Guidelines for Electronic Text Encoding and Interchange*. [6.4.2015]. http://www.tei-c.org/P5/ .

Wagner, Wiltrud (in Vorbereitung): *Idioms and Ambiguity in Context: Compositional and Phrasal Readings of Idiomatic Expressions*. Dissertation. Tübingen.

Winkler, Susanne (eds.) (2015): *Ambiguity: Language and Communication*. Berlin: de Gruyter.

Winter-Froemel, Esme / Zirker, Angelika (2010): „Ambiguität in der Sprecher-Hörer-Interaktion. Linguistische und literaturwissenschaftliche Perspektiven“, in: Klein, Wolfgang / Winkler, Susanne (eds.): *Ambiguität*. Zeitschrift für Literaturwissenschaft und Linguistik 40 (158). Stuttgart: Metzler 76–97.

Winter-Froemel, Esme / Zirker, Angelika (2015): „Ambiguity in Speaker-Hearer-Interaction: A Parameter-Based Model of Analysis“, in: Winkler, Susanne (eds.): *Ambiguity: Language and communication*. Berlin: de Gruyter 283–339.

Analyzing Features for the Detection of Happy Endings in German Novels

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

Reger, Isabella
isabella.reger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Zehe, Albin
zehe@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Becker, Martin
becker@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Hettinger, Lena
lena.hettinger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Hotho, Andreas
hotho@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Note: An English version of this paper is available from https://arxiv.org/abs/1611.09028 .

Der Plot ist ein grundlegendes Strukturelement literarischer Texte. Dementsprechend wären Methoden zur computergestützten Repräsentation von Plot oder bestimmten Plot-Elementen ein großer Gewinn für die quantitative Literaturanalyse. Dieses Paper betrachtet ein solches Plot-Element: das Ende; genauer gesagt untersuchen wir die Frage, ob ein Werk ein Happy End hat oder nicht. Dazu setzen wir Sentimentanalyse ein, wobei wir den Fokus auf die qualitative Betrachtung bestimmter Features und deren Performanz legen, um tiefere Einsicht in die Funktionsweise der automatischen Klassifikation zu erhalten. Außerdem zeigen wir, wie die beschriebene Vorgehensweise auf nachfolgende Forschungsfragen angewendet werden und dabei zu interessanten Ergebnissen hinsichtlich der Erscheinungszeit der Romane führen kann.

Verwandte Arbeiten

In einer der ersten Arbeiten beschäftigt sich Mark Finlayson mit folkloristischen Erzählungen und entwickelt einen Algorithmus, der Ereignisse erkennt und daraus übergeordnete Konzepte wie Niedertracht oder Belohnung abstrahiert (Finlayson 2012). Reiter et al. identifizieren Ereignisse sowie deren Teilnehmer und Reihenfolge und nutzen maschinelle Lernverfahren, um strukturelle Ähnlichkeiten über Erzählungen hinweg aufzudecken (Reiter 2013, Reiter et al. 2014).

In letzter Zeit richtet sich einige Aufmerksamkeit auf die Sentimentanalyse, insbesondere seit Matthew Jockers emotionale Erregung als Indikator für Plotstrukturen vorgeschlagen hat (Jockers 2014). Er unterteilt Romane in Segmente und bildet daraus emotionale Plot-Kurven (Jockers 2015). Obwohl die Idee, Sentimentanalyse in diesem Zusammenhang einzusetzen, gut aufgenommen wurde, wurde Jockers für seine Verwendung der Fourier-Transformation zur Glättung der resultierenden Plot-Kurven kritisiert (Swafford 2015, Schmidt 2015).

Micha Elsner (Elsner 2015) verwendet, neben anderen Features, ebenfalls Sentimentkurven, um Repräsentationen des Plots romantischer Werke zu erstellen. Er verknüpft diese Kurven mit bestimmten Figuren und untersucht auch das gemeinsame Auftreten von Figuren. Die Auswertung seines Ansatzes zeigt, dass er echte Romane mit beachtlichem Erfolg von künstlich umgestellten Versionen unterscheiden kann, was darauf hindeutet, dass seine Methoden tatsächlich bestimmte Aspekte der Plotstruktur abbilden.

In vorhergehenden Arbeiten haben wir Sentiment-Features verwendet, um Happy Ends, als ein wichtiges Plot-Element, in deutschsprachigen Romanen zu erkennen, wobei wir einen F1-score von 73% erreichen konnten (Zehe et al. 2016).

Korpus und Ressourcen

Unser Datensatz besteht aus 212 deutschsprachigen Romanen, die hauptsächlich aus dem 19. Jahrhundert stammen. [1] Zu jedem Roman wurde manuell annotiert, ob er ein Happy End hat (50%) oder nicht (50%). Die dafür relevanten Informationen stammen aus den Zusammenfassungen des Kindler Literatur Lexikon Online [2] und aus Wikipedia [3] . Sofern keine Zusammenfassung eines Romans verfügbar war, wurde das Ende von den Annotatoren gelesen.

Unsere Sentimentanalyse erfordert eine Ressource, die auflistet, welche Gefühle Leser typischerweise mit bestimmten Worten oder Phrasen eines Textes assoziieren. Dieses Paper verwendet das NRC Sentiment Lexikon (Mohammad und Turney 2013), zu dem eine automatisch übersetzte deutsche Version verfügbar ist [4] . Eine besondere Eigenschaft dieses Lexikons ist, dass zu jedem Wort neben je einem binären Wert (0 oder 1) für positive und negative Konnotation (2 Features) auch seine Zugehörigkeit zu 8 Basisemotionen (Wut, Angst, Ekel, Überraschung, Freude, Vorfreude, Vertrauen und Trauer) festgehalten ist (vgl. Tabelle 1). Zusätzlich ermitteln wir die Polarität eines Wortes, indem der negative vom positiven Wert abgezogen wird (ein Wort mit einem positiven Wert von 0 und einem negativen Wert von 1 erhält also die Polarität -1). Die Polarität dient als ein zusammengefasster Emotionswert. Insgesamt betrachten wir also 11 Features.

Tabelle 1: Beispieleinträge aus dem NRC Sentiment Lexikon

Wort/ Dimension	verabscheuen	bewundernswert	Zufall
Positiv	0	1	0
Negativ	1	0	0
Polarität	-1	1	0
Wut	1	0	0
Vorfreude	0	0	0
Ekel	1	0	0
Angst	1	0	0
Freude	0	1	0
Trauer	0	0	0
Überraschung	0	0	1
Vertrauen	0	1	0

Experimente

Ziel dieses Papers ist es, Features, die zur Erkennung von Happy Ends in Romanen genutzt wurden, genauer zu untersuchen, um Einsichten in die Relevanz bestimmter Features zu erhalten. Dazu übernehmen wir die Features und Methoden, wie sie in Zehe et al. (2016) beschrieben sind. Die Parameter der linearen SVM sowie die Einteilung in 75 Segmente sind ebenfalls aus diesem Paper übernommen.

Features. Da keine verlässlichen Kapitelannotationen verfügbar waren, wurde jeder Roman in 75 gleichgroße Blöcke unterteilt, die wir als *Segmente* bezeichnen. Für jedes lemmatisierte Wort werden die oben beschriebenen 11 Sentiment-Werte ermittelt. Anschließend wird für jedes Segment der entsprechende Durchschnitt berechnet, sodass 11 Werte pro Segment vorliegen. Diese werden als ein Feature-Set betrachtet.

Qualitative Feature-Analyse. Da unser Korpus zu gleichen Teilen aus Romanen mit und ohne Happy End besteht, erreichen sowohl die Random Baseline, als auch die Mehrheits-Baseline eine Klassifikationsgenauigkeit von 50%.

Aufgrund unserer Annahme, dass die relevante Information zur Klassifikation von Happy Ends am Ende eines Romans zu finden ist, wurden zunächst die Sentiment-Werte des letzten Segments als einziges Feature-Set ($f_{d,\ n}$) verwendet, was zu einer Genauigkeit von 67% führte.

Um unserer Intuition gerecht zu werden, dass nicht nur das letzte Segment an sich, sondern auch sein Verhältnis zum Rest des Romans für die Klassifikation von Bedeutung ist, wurden sogenannte Sektionen (*sections*) eingeführt: das letzte Segment eines Romans bildet die *final section*, während die übrigen Segmente zur *main section* gehören. Über die Sektionen wurden wiederum Durchschnittswerte gebildet, indem der jeweilige Wert aller 11 Features über alle Segmente in der betreffenden Sektion gemittelt wurde. Um das Verhältnis zwischen diesen Sektionen abzubilden, wurden die Differenzen zwischen den Sentiment-Werten der final section und den durchschnittlichen Sentiment-Werten aller Segmente in der main section als zusätzliche Features betrachtet. Dies hatte jedoch keine Auswirkungen auf die Ergebnisse.

Diese Beobachtung führte uns zu der Annahme, dass unser Begriff des "Endes" nicht differenziert genug ist, da die Anzahl an Segmenten für jeden Roman und damit auch die Grenzen des finalen Segments relativ willkürlich gewählt wurden. Daher wurde die Aufteilung in final section und main section im Folgenden variiert, sodass die final section mehr als nur das letzte Segment enthalten kann.

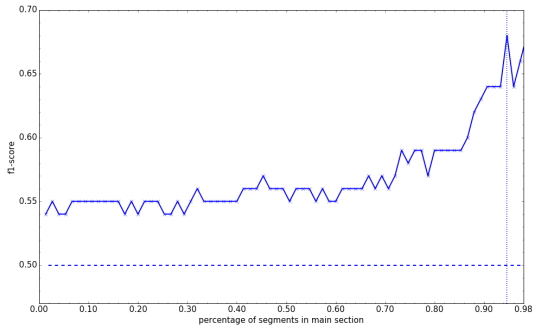


Abbildung 1: Klassifikationsgenauigkeit für verschiedene Unterteilungen in main und final section. Die gestrichelte Linie gibt die Baseline an, die gepunktete Linie markiert die Aufteilung, bei der der maximale F1-score erreicht wird.

Abbildung 1 zeigt, dass die Klassifikationsgenauigkeit steigt, wenn mindestens 75% der Segmente in der main section sind und ein Maximum bei ca. 95% erreicht (bei 75 Segmenten insgesamt bedeutet das 4 Segmente in der final section und 71 Segmente in der main section). Mit dieser Aufteilung verbessert sich der F1-Wert auf 68%, wenn nur das Feature-Set der final section ($f_{d,\ final}$) verwendet wird, und weiter auf 69%, wenn die Differenzen zu den durchschnittlichen Sentiment-Werten der main section ($f_{d,\ main\ -\ final}$) miteinbezogen werden.

Da sich die Ergebnisse durch die Einbeziehung des Verhältnisses zwischen der final section und der main section verbessert

haben, war unser nächster Schritt, den Verlauf der Sentimentkurve gegen Ende eines Romans genauer zu modellieren. Beispielsweise könnte sich kurz vor dem Ende eine Katastrophe ereignen, die anschließend im Sinne eines Happy Ends aufgelöst wird. Um diese Intuition abzubilden, führten wir eine weitere Sektion ein, die sogenannte late-main section, die die letzten Segmente der main section umfasst. Die Differenzen zwischen den Feature-Sets für die late-main section und die final section wurden als zusätzliche Merkmale verwendet ($f_{d,\ late\text{-}final}$). Mit diesen 3 Feature-Sets erzielten wir einen F1-score von 70%. Durch die zusätzliche Verwendung des Feature-Sets für das letzte Segment stieg der F1-score auf 73%.

Tabelle 2: F1-score für die verschiedenen Feature-Sets

Features	Ergebnisse
1) Feature-Set finales Segment	67%
2) Feature-Set finales Segment und Differenz zur main section	67%
3) Feature-Set final section mit final section der Länge 4	68%
4) Feature-Set 3 und Differenz zur main section	69%
5) Feature-Set 4 und Differenz zwischen late-main section und final section	70%
6) Feature-Set 5 und Feature-Set finales Segment	73%

Die beschriebenen Ergebnisse sind in Tabelle 2 zusammengefasst. Hier wird deutlich, dass die Aufnahme der einzelnen Feature-Sets jeweils zu einer kleinen Verbesserung geführt hat, bis hin zu einem F1-score von 73%. Obwohl die Aufteilung mit 4 Segmenten in der final section die besten Ergebnisse erzielte, konnten wir auch beobachten, dass einige Romane mit mehreren verschiedenen Unterteilungen korrekt klassifiziert werden konnten. Andere Romane hingegen konnten in keinem Setting korrekt vorhergesagt werden. Als Beispiel sei hier Jules Vernes Roman *Zwanzigtausend Meilen unter dem Meer* genannt, der ein eindeutiges Happy End mit klaren Grenzen hat, das jedoch extrem kurz ist und nur aus den ca. 250 letzten Wörtern besteht. Diese Beobachtungen zeigen, dass der Begriff des "Endes" eines Roman sehr variabel ist und von Text zu Text sehr unterschiedlich manifestiert sein kann.

Korrelation mit Erscheinungszeit. Das wirft wiederum die Frage auf, ob die Sensibilität unserer Methode hinsichtlich solcher Variabilität genutzt werden kann, um gewisse Eigenschaften der Romane in unserem Korpus besser zu verstehen. Als Beispiel haben wir untersucht, ob und inwiefern der Erfolg verschiedener Unterteilungen von Romanen mit deren Erscheinungsdatum korrelieren. Um die Ergebnisse so gut wie möglich interpretierbar zu halten, beschränken wir uns auf ein Feature-Set: die Sentiment-Werte der finalen Sektion. Zunächst haben wir unser Korpus in 4 Gruppen unterteilt: Romane, die vor 1830 erschienen sind (65 Texte), zwischen 1831 und 1848 (31 Texte), zwischen 1849 und 1870 (29 Texte) und nach 1871 (87 Texte). Diese Einteilung führte zu ähnlich großen Untergruppen, von denen keine eine besondere Tendenz hinsichtlich Romanen mit oder ohne Happy End aufweist.

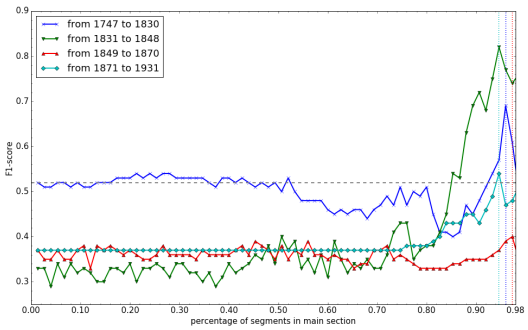


Abbildung 2: F1-score für verschiedene Unterteilungen in main und final section. Die farbigen Kurven stehen für Romane aus verschiedenen Zeitperioden. Die gestrichelte Linie zeigt die Zufallsbaseline für die Zeitperiode ab 1871. Die Baselines für die anderen Zeitperioden liegen etwas darunter und werden daher nicht dargestellt. Die gepunkteten Linien zeigen jeweils den maximalen F1-Wert für die entsprechende Zeitperiode.

Abbildung 2 zeigt, dass die Klassifikation erneut dann am besten funktioniert, wenn ca. 95-98% der Segmente in der Hauptsektion sind, unabhängig von der Zeitperiode. Die beste Aufteilung in Sektionen korreliert also nicht mit dem Erscheinungsjahr eines Romans. Es fällt jedoch auf, dass die Romane nach 1848 deutlich niedrigere Werte liefern als die vor diesem Jahr veröffentlichten Texte, meistens sogar unterhalb der Baseline. Das deutet auf eine Korrelation zwischen dem Erscheinungsdatum

und der Klassifikationsgenauigkeit hin: Vor dem Realismus erschienene Romane sind hinsichtlich des Happy Ends leichter zu klassifizieren als realistische Romane. Eine mögliche Erklärung für diese Beobachtung könnte die stärker schematische Struktur der vor-realistischen Romane sein.

Wir sind uns bewusst, dass die Anzahl der Romane für die einzelnen Zeitperioden relativ klein ist, sodass diese Beobachtungen zunächst als exploratorische Einblicke gesehen werden müssen. Nichtsdestotrotz zeigen diese vorläufigen Ergebnisse, dass die automatische Erkennung von Happy Ends, sogar mit nur einem recht einfachen Feature-Set, Zusammenhänge zu anderen Eigenschaften von Romanen aufdecken kann, die für die Literaturwissenschaft von großem Interesse sind.

Fazit und zukünftige Arbeiten

Die automatische Erkennung von Happy Ends als wesentlichem Plot-Element von Romanen ist ein nützlicher Schritt in Richtung einer umfassenden computergestützten Repräsentation des Plots literarischer Texte. Unsere Experimente zeigen, dass verschiedene Features auf Basis von Sentimentanalyse eine Erkennung von Happy Ends in Romanen mit unterschiedlicher, aber insgesamt solider Genauigkeit ermöglichen. Obwohl unser Ansatz relativ einfach gehalten ist, kann er zu substantiellen Erkenntnissen für die Literaturwissenschaft führen.

In zukünftigen Arbeiten soll die Genauigkeit unserer Methode verbessert werden, indem die hohe Variabilität des Endes in Romanen differenzierter betrachtet wird. Außerdem könnte der Ansatz eingesetzt werden, um bestimmte Eigenschaften weiterer Romankorpora tiefergehend zu untersuchen.

Fußnoten

1. Quelle: https://textgrid.de/digitale-bibliothek
2. www.kll-online.de
3. https://de.wikipedia.org
4. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Bibliography

Elsner, Micha (2015): „Abstract Representations of Plot Structure“, in: *Linguistic Issues in Language Technology* 12 (5).

Finlayson, Mark A. (2012): *Learning Narrative Structure from Annotated Folktales*. PhD thesis, Massachusetts Institute of Technology.

Jockers, Matthew L. (2014): *A novel method for detecting plot*. http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/ [letzter Zugriff 25. August 2016].

Jockers, Matthew L. (2015): *The rest of the story*. http://www.matthewjockers.net/2015/02/25/the-rest-of-the-story/ [letzter Zugriff 25. August 2016].

Mohammad, Saif / Turney, Peter (2013): „Crowdsourcing a Word-Emotion Association Lexicon“, in: *Computational Intelligence* 29 (3): 436–465.

Reiter, Nils (2013): *Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms*. PhD thesis, Heidelberg University.

Reiter, Nils / Frank, Anette / Hellwig, Oliver (2014): „An NLP-based Cross-Document Approach to Narrative Structure Discovery“, in: *Literary and Linguistic Computing* 29 (4): 583–605 10.1093/llc/fqu055.

Schmidt, Benjamin M. (2015): *Commodius vici of recirculation: the real problem with Syuzhet*. http://benschmidt.org/2015/04/03/commodius-vici-of-recirculation-the-real-problem-with-syuzhet/ [letzter Zugriff 25. August 2016].

Swafford, Annie (2015): “ „Problems with the Syuzhet Package“. https://annieswafford.wordpress.com/2015/03/02/syuzhet/ [letzter Zugriff 25. August 2016].

Zehe, Albin / Becker, Martin / Hettinger, Lena / Hotho, Andreas / Reger, Isabella / Jannidis, Fotis (2016): „Prediction of Happy Endings in German Novels“, in: *Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing 2016*.

Anybody out there? Der Begriff der Masse im Crowdsourcing

Schilz, Andrea
andrea.schilz@uni-passau.de
Universität Passau, Deutschland

Thematik

Die „Masse“ als „Beschaffer“ definierter Informationen hat sich ihren Platz im Feld der Digital Humanities bzw. in der Schnittmenge von DH und Institutionen für den Erhalt kulturellen Erbes erobert. Aufgaben ganz verschiedener Schwierigkeitsgrade werden an eine anonyme Menge interessierter Individuen delegiert, die so zu produktiven Beiträgern spezifischer Projekte werden. Oft zitierte Beispiele sind Old Weather und Transcribe Bentham (Dunn/Hedges 2012), beides Transkriptions-Projekte; auf dem Gebiet von OCR-Korrektur und Annotation wird auch mit Gamification gearbeitet (Digitalkoot, Metadata Games, ARTigo).

Das Thema Crowdsourcing zeigt sich in den DH freilich nicht nur empirisch; es liegen Untersuchungen zu Typologien und Modellen sowie Fallstudien vor, und auch ethische Fragen werden verhandelt. Der Diskurs tendiert dabei dazu, das im kommerziellen Crowdsourcing zentrale Argument des Ökonomischen als eher sekundär einzustufen. Vorrangig wird dem crowdbasierten Generieren wissenschaftlich verwertbarer Daten bzw. dem nachhaltigen Anreichern digitalisierter Quellen hinsichtlich des weichen Faktors einer Sensibilisierung für kulturelles Erbe hohes Potential zugesprochen. Dabei gibt es für das zielführende Einsetzen demokratisierten Wissens eine Blaupause, die naturwissenschaftlich ausgerichtete Citizen Science. Die Plattform Zooniverse spielt hierbei eine namhafte Rolle, obgleich auch analoge Varianten nach wie vor ihre Berechtigung haben (bekannt sind die saisonalen ornithologischen Zählungen).

Ausgehend von dieser Gesamtsituation entwickelt dieser Beitrag Fragenkomplexe zu einer kulturwissenschaftlichen Kontextualisierung des Crowdsourcings für die DH. Die Kernfrage ist dabei, inwiefern sich geisteswissenschaftliche Konnotationen des Masse-Begriffs in Konzepten zum „Gebrauch“ der Crowd spiegeln und welche gesellschaftliche Relevanz daraus erwächst.
Angrenzend wird betrachtet, wie sich Crowdsourcing vor der Digitalität gestaltete und welche „Crowd“ gemeint war. Außerdem:Wo sind Chancen und Grenzen des Crowdsourcings zu detektieren, bezüglich Machbarkeit, Effizienz, Ethik, und welche Schlüsse ziehen die DH daraus? Was lässt sich aus der Situation für den Anspruch an eine nachhaltige Datengenerierung ableiten?

Begründung

Zwei Determinanten bestimmen die Themenwahl. Erstens: Im Feld der Digital Humanities ist Crowdsourcing ein Thema. Zweitens: Kernfaktor des Crowdsourcings ist die Ressource „Masse“.

Aktuelle Positionen in den DH begreifen, wie oben skizziert, Crowdsourcing nicht als bloßes Produktionsmittel. Mein Beitrag baut auf diesen Untersuchungen auf und ergänzt sie durch die These, dass für die DH zu einem adäquaten Umgang mit der Methode Crowdsourcing auch eine kritisch angebundene Diskussion des Masse-Begriffs unter dezidiert kultur- bzw. geisteswissenschaftlichen Gesichtspunkten gehören sollte. Ich nähere mich dem Phänomen Crowdsourcing unter diesen Prämissen an, unter Deklaration einer notwendigen Subjektivität: Meine Argumentationslinie stellt eine von vielen Optionen dar – andere mögen weitere hinzufügen.

Prolog: Analoges Crowdsourcing

Das Delegieren definierter Aufgaben an potentiell unbekannte Zuträger ist nicht erst ein Phänomen des digitalen Zeitalters. Es gab in der Geschichte der Geistes- und Kulturwissenschaften einige prominente Projekte mit akkumulativem Charakter, die Beiträger (aus einem vorher definierten Pool) über Aufrufe akquirierten und auf aktive positive Reaktion entscheidend angewiesen waren.
Beispiele sind die Aufrufe vonGeorge Perkins Marsh für das *New English Dictionary* 1859 (Ridge), die erste groß angelegte volkskundliche Fragebogenaktion von Wilhelm Mannhardt 1865 zu „alten agrarischen Gebräuche(n) und Erntesitten“ sowie das Kulturraum-

Forschungsprojekt *Atlas der deutschen Volkskunde* (1930-1935).

Kritik der Masse

Die Masse umgibt eine semantische Aura, die ganz bestimmten Bildern entstammt und in diese mündet: Amorph, wesenhaft, unberechenbar, entindividualisiert, und in dialektischer Weise lenkbar und unsteuerbar. Die Masse als Negativ zur Selbstbestimmtheit ist, naheliegend, historisch unterfüttert, nach Links wie nach Rechts. Das Zurückweisen der Masse als Identifikationsmoment wird, auch dies liegt nahe, bestimmt durch eine von räumlichen, zeitlichen und soziopolitischen Rahmenbedingungen bedingte Enkulturation. In Anlehnung an Sartre ließe sich formulieren: „Die Masse, das sind die anderen." Dieser Abwehrreflex weist eine kulturgeschichtliche Dimension auf, die im Folgenden in groben Zügen freigelegt wird.

Die gefährliche und die dumme Masse

Gustave Le Bon, früher und wirkmächtiger Vertreter der Massenpsychologie, sah „die Massen" als kulturzerstörende Kraft, wobei der Kulturbegriff, zeittypisch, mit dem Moment der „Rasse" gekoppelt wird. In Le Bons Konzept der Masse wird, wiederum zeittypisch, ein Deutungsmuster manifest, das Eingang ins kulturelle Gedächtnis finden wird: Masse/Tief vs. Kultur/Hoch.

Ein Misstrauen in die Masse und eine daraus resultierende Angst vor ihrer nivellierenden Macht fand insbesondere in der Figur der Massenseele Ausdruck. Der so abstrahierten Masse werden quasi metaphysische Eigenschaften zugewiesen, woraus sich (unschwer) ein Negieren des Individuums ableiten lässt. Exemplifiziert wird dies anhand einer kurzen Skizze zum Überbevölkerungsdiskurs, der in den 1970er Jahren global erneuert wurde und die historisch gesetzte Verwerfung Arm/Reich um jene von Norden/Süden erweitert hat.

Komplementär zur Gefährlichkeit ist die Dummheit, welche der Masse distinktiv zugeschrieben wird. Dies hat in der Geisteswissenschaft früh ein Echo erzeugt, das in dekonstruierende Reaktion geht. Bei Marx und Engels werden Ansätze erkennbar, den Masse-Begriff als elitäres Produkt herauszuarbeiten. Zeitgenosse Charles Mackay formulierte mit der provokanten These vom „Wahnsinn der Massen" eine Kritik sozialer Mechanismen. Gut hundert Jahre später definiert Pierre Bourdieu Muster der Distinktion in Abgrenzung zur Masse und präzisiert das Affirmative des Nicht-Massenmenschen.

Die unheimliche und die konstruktive Masse

Zoomt man in der Analyse der Masse von der Makroebene zur Mikroebene heran, begegnet man dem „Massenmenschen". Berechtigtes Misstrauen in ihn äußert beispielsweise Walter Benjamin. Überhaupt steht die Frankfurter Schule der Masse bzw. den für sie bestimmten Produkten bekanntlich skeptisch gegenüber. An diesem Punkt der Erzählung von der „modernen Masse" stoßen wir aber auch auf eine parallele Lesart, mit der ein ästhetisches Fassen des Massenmenschen einhergeht, das auf das Wesen der Popkultur verweisen wird. Poes ungreifbarer, geschichtsloser „Man in the Crowd" scheint auf, sowie Baudelaires „Heimat" im Urbanen, in der die Singularisierung als Chance für ein Man-selbst-sein in der Menge begriffen wird.

Benjamin deutet, Bezug nehmend auf Baudelaires Auffassung vom Flaneur in der Menge, diese Figur kritisch als Metapher für die Gefahr des Scheins, der vom Kollektiv ausgeht. Doch der Flaneur bietet auch die Möglichkeit einer positiven Umdeutung vor dem Hintergrund des Gemeinnützigkeit-Gedankens. Es geht dabei um das Gebiet der Wissensallmende: um gemeinsames Gut der Informationsgesellschaft – digitales Gemeingut. Wikipedia übersetzt „Allmendefertigung durch Gleichberechtigte" für Commons-based Peer Production (CBPP), ein von Yochai Benkler geprägter Begriff. Er führt uns zur Masse im digitalen Raum bzw. zum Flanieren im WWW, dem Browsen. Mit welchen Konzepten begegnen wir potentiellen Partizipatoren unserer Crowd?

Konzepte

Wer die „Weisheit der Masse" (James Surowiecki) nutzt, setzt auf Schwarmintelligenz: Kollektive Lösungen sind besser als individuelle, lautet das Credo. Unter welchen ökonomischen, methodischen und ethischen Implikationen ein für das eigene Ziel nützlicher Schwarm generiert wird, unterscheidet sich jedoch maßgeblich.

Task und Methode

Das ökonomisch boomende Modell des Microtaskings kommt dem am nächsten, was Jeff Howe, der den Begriff Crowdsourcing prägte, als „Future of Business“ (Howe) sieht. Materielle Basis sind diskrete, einfache Aufgaben, die, in Arbeitspakete gesplittet, effizienter durch Menschen - Crowdworker - als mit Maschinen gelöst werden. Prominent ist Amazon's Mechanical Turk, doch auch in den DH wird mit kommerziell basiertem Crowdwork experimentiert.

Microtasking findet sich jedoch auch in ganz anderer Weise in DH-Projekten wieder, die gemäß dem CBPP-Ansatz von der Grundannahme einer reflektiert agierenden Masse, konstituiert über selbstbestimmte Individuen, getragen werden. Die Methode der Folksonomy ist hier zu nennen, auf die z. B. die gamifizierte Metadaten-Sammlung ARTigo setzt sowie andere niederschwellige, spielerische Settings, angeboten als Open Source über die Plattform Metadata-Games.

Methodisch angrenzend an das Microtasking findet sich das Macrotasking (Brandon Walsh). Bei diesem Crowdsourcing-Konzept steht Akkumulation statt Aggregation im Vordergrund, neue komplexe Informationen werden generiert – flächig bekanntes Beispiel ist hier Wikipedia. Macrotasking ist skizzierbar als Microtasking mit Spezialisierungsbedarf seitens der Crowd, im Feld der DH ein häufiges Desideratsprofil. Etwa auf dem Feld des Transkribierens: Exemplarisch wird aufgezeigt, wie variantenreich Konzepte, Vorgehen und Bereiche sich darstellen können, und was dies für Auswirkungen auf Projektspezifikationen hat. Aspekte sind dabei Formatierung, Annotation sowie das strukturierte Akkumulieren verteilter Daten (Kearney/Wallis).

Motiv und Modell

Für das ergebnisorientierte Akquirieren einer Crowd bedarf es einerseits gezielter Kommunikation im Vorfeld – der Aufruf ist ein wesentliches Erfolgsmoment. Was motiviert andererseits dazu, Angebote anzunehmen und, nicht weniger relevant, sie mittel- bis längerfristig verlässlich teilnehmend zu verfolgen? Grundlegende Argumente sind Bezahlung, Bindung, Teilen und Spielen (Oomen/Aroyo). Zudem müssen die Beschaffenheit der Quellen, Ergebnis-Desiderate und Konzepte fein abgestimmt werden, um Projekte erfolgreich zu realisieren.

In den DH wurden differenzierte Typologien entwickelt, um insbesondere auch das Problem der Akquise methodologisch zu schärfen. Flankiert wird dies durch eine Palette an Open Source-Werkzeugen, die niederschwellige Einstiegsoptionen durch benutzerfreundliche Schnittstellen und klare Strukturen schaffen. DH-geeignete Projektdeterminanten sind bereits erprobt worden, empirische Erfahrungen indizieren jedoch öfters ein asymmetrisches Verhältnis im Profil der liefernden Crowd: Es sind nur wenige Beiträger, die den Hauptanteil an der Bearbeitung tragen.

Nachhaltigkeit

Ein Kernargument im DH-Diskurs für das zielführende Erzeugen und Binden einer Crowd ist das Moment der Identifikation – ein Teil von etwas Großem zu sein (Terras). Was sagt dies aus über die Masse und über jene, die sie nutzen wollen? In Bezug auf angesprochene Konnotationen der Masse wird kursorisch diskutiert, welche Crowd wir wollen: Die kontrollierbare und effiziente, die die Arbeit erledigt, oder die interessierte und empathische, die um Kulturerhalt besorgt ist. Ist beides möglich? An dieser Stelle wird bilanziert, wo Möglichkeiten und Grenzen des Crowdsourcings in den DH liegen und welche Konsequenzen daraus in puncto Nachhaltigkeit zu benennen sind – nicht nur bezüglich Datengenerierung und -optimierung, sondern auch explizit hinsichtlich des Faktors Mensch.

Archival Cultural Heritage Online: Eine Virtuelle Forschungsumgebung im Spannungsfeld von Open Access, Nachhaltigkeit und Datenschutz

Lange, Felix
flange@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Berlin

Wintergrün, Dirk
dwinter@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Berlin

Wannenwetsch, Oliver
oliver.schmitt@gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mbH, Göttingen

Schoepflin, Urs
schoepfl@mpiwg-berlin.mpg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mbH, Göttingen

Wie sich die langfristige wissenschaftliche Nutzbarkeit von großen digitalen Datenrepositorien sicherstellen lässt, ist eine in den letzten Jahren in der DH-Community und darüber hinaus intensiv diskutierte und noch nicht abschließend geklärte Frage. [1] In den Digitalen Geisteswissenschaften werden in diesem Zusammenhang zur Zeit vorrangig Probleme der technischen Nachhaltigkeit und der Datenstandards diskutiert [Fornaro (2016)]. Im Hinblick auf Repositorien für die gegenwartsnah arbeitenden geistes- und sozialwissenschaftlichen Disziplinen wie die Zeitgeschichte sind aber auch komplexe datenschutz- und urheberrechtliche Anforderungen zu berücksichtigen. Eine im Archivwesen diskutierte Antwort auf diese Herausforderung ist es, die rechtliche Absicherung des Zugangs zu Digitalisaten in sog. „digitalen Lesesälen“ zu organisieren, die einen Zugriff ausschließlich in den Räumlichkeiten des jeweiligen Archivs zulassen [Plassmann (2016), S. 219]. Dabei wird aber das Ziel der Open-Access-Bewegung, wissenschaftliche Quellen und Forschungsergebnisse einer möglichst großen Fachöffentlichkeit zugänglich zu machen, verfehlt. [2] In den Sozialwissenschaften hat die Brisanz dieser Frage bereits zur Gründung von Datenzentren geführt, die Fragen der technischen *und* der rechtlichen Datensicherheit in den Mittelpunkt stellen. [3] Der vorliegende Beitrag stellt mit Archival Heritage Online – ArCHO eine digitale Forschungsinfrastruktur vor, die dazu dient, das Verlangen nach offener wissenschaftlicher Nutzung mit den rechtlichen Bedingungen für den nachhaltigen Zugang zu zeitgeschichtlichem Archivmaterial in Einklang zu bringen. [4]

Der prototypische Anwendungsfall für ArCHO ist das seit 2014 laufende und auf zunächst fünf Jahre angelegte Forschungsvorhaben „Geschichte der Max-Planck-Gesellschaft“ (GMPG). [5] Es untersucht die Geschichte der MPG von ihrer Gründung im Jahre 1948 bis zum Jahr 2002 und legt dabei den Schwerpunkt auf institutsübergreifende Fragestellungen zu Themenfeldern wie Periodisierungen, Innovationen, Internationalisierung, Forschung und Wirtschaft, Gender und Wissenschaft sowie Konkurrenz und Kooperation. Diese Themen lassen sich naturgemäß nicht allein durch kleinere Fallstudien bearbeiten, sondern erfordern thematisch und chronologisch breit angelegte Querschnittsuntersuchungen mit einer entsprechend umfänglichen Quellengrundlage. Aus diesem Grund wird im Laufe des Projektes ein großes digitales Textkorpus angelegt, dessen Schwerpunkt Digitalisate von mehreren Regalkilometern an Verwaltungsschriftgut aus der Generalverwaltung der MPG und einzelnen Instituten bilden. Desweiteren werden thematisch spezialisierte Datenbestände wie eine Patent- und eine Personendatenbank sowie ein digitales Korpus mit Veröffentlichungen der MPG aufgebaut. Mit dafür entwickelten oder angepassten Tools [Kruse et al. (2015)] lassen sich so beispielsweise Konjunkturen von Forschungsthemen, unterschiedliche professionelle Netzwerke zwischen WissenschaftlerInnen und wissenschaftliche Karrierewege erforschen. Im Sinne der guten wissenschaftlichen Praxis [6] sollen die Arbeitsergebnisse, also sowohl die

digitalisierten und annotierten Quellen als auch alle statistischen Auswertungen, mindestens zehn Jahre nach Projektende abrufbar bleiben. ArCHO als digitales Findmittel und Analyseplattform ist daher mit einem Fokus auf langfristiger Verfügbarkeit von Forschungsdaten konzipiert worden. Dabei wurde eine Nachhaltigkeitsstrategie entwickelt, die der noch ungeklärten Aufgabenteilung zwischen Forschungseinrichtungen, Gedächtnisinstitutionen sowie Daten- und Rechenzentren bei der Langzeitarchivierung geisteswissenschaftlicher Forschungsdaten Rechnung trägt. Denn diese Aufgabe kann angesichts der großen technischen Komplexität und des Wartungsaufwandes für Virtuelle Forschungsinfrastrukturen sowie der großen Menge an vorzuhaltenden Daten nicht allein Gedächtnisinstitutionen wie wissenschaftlichen Archiven überantwortet werden. Andererseits sind Rechen- und Datenzentren nur bedingt dazu in der Lage, neben dem Archivrecht auch komplexe spezifische Zugangsregeln für einzelne Datenrepositorien umzusetzen. Daher ermöglicht es ArCHO mit einem verlässlichen Zugangsmanagement, dass die Forschungseinrichtung den Zugang selbst rechtssicher regeln kann.

Die in ArCHO implementierte Zugangsverwaltung setzt auf eine starke Differenzierung von Nutzerrollen einerseits und von Bestandteilen einzelner Datensätze andererseits. Auf der Nutzerseite muss beispielsweise im Anwendungsfall GMPG unterschieden werden zwischen der wissenschaftlichen Öffentlichkeit, Forschern innerhalb des Forschungsvorhabens mit einem privilegierten Zugang zu den Aktenbeständen und einem Projektkollegium, das besonders sensible Datenbestände nach einer Einzelfallprüfung für die Forscher freigibt. Weitere Abstufungen von Zugangsrechten können sich aus spezifischen Aufgabenbereichen bei der Dateneingabe und -verwaltung ergeben [vgl. Neuroth et al. (2010), 16:14 ff.]. Die Aufgabe der Zugangsregelung muss auch nach Projektende weiter von dazu befugten Personen ausgeübt werden können und ist daher eine wichtiger Nachhaltigkeitsaspekt. Denn beispielsweise ist bei datenschutzrechtlich sensiblen Dokumenten mit Personenbezug, deren Sichtung durch Forscher der Einwilligung der betroffenen Personen bedarf, je nach konkreter rechtlicher Ausgestaltung diese Bewilligung an das Forschungsvorhaben und damit an dessen Laufzeit gebunden. Die Nutzungserlaubnis erlischt in diesen Fällen nach Projektende und entsprechend muss auch der digitale Zugang verwehrt werden. Auf der anderen Seite werden manche Akten erst nach Ende der Archivschutzfrist vollständig nutzbar, was in einer nachhaltigen Forschungsinfrastruktur ebenfalls berücksichtigt werden sollte.

Auf der Datenseite ermöglicht ArCHO eine starke Differenzierung von einzelnen zu einem Dokument gehörenden Daten mit dem Ziel, unter Einhaltung der rechtlichen Vorgaben möglichst viele Informationen für die Forschung zur Verfügung zu stellen. So sind bei einer Personalakte mit sensiblen Inhalten möglicherweise die Signatur, Laufzeit und Angaben zur inhaltlichen Klassifikation durch das haltende Archiv nicht schutzwürdig, wohl aber der Volltext und der Titel. Es kann also je nach Bestand jedes Metadatum und jedes Derivat des Digitalisates (OCR-Erfassungen u.a.) eine andere Schutzwürdigkeit haben. Die Gesamtzahl dieser Regeln, die zwischen beliebigen Typen von (Meta-)Daten unterscheiden, und die Vielzahl von abgestuften Nutzerrechten führen zu einer Matrix aus Nutzerrollen und Teildatensätzen, deren einzelne Werte sich stets ändern können. Sie wird technisch realisiert durch einen sogenannten *Policy Decision Point* (PDP). Dabei handelt es sich um ein außerhalb des eigentlichen Dokumentkorpus angesiedeltes und technisch eigenständiges Softwaremodul, das zwischen der Nutzer-Datenbank und dem Korpus vermittelt.

Die Umsetzung eines solchen Rechtemodells innerhalb einer ansonsten marktüblichen Webanwendung leistet den oben geschilderten Anforderungen aber noch nicht Genüge. Denn ein solches System wäre höchst verwundbar gegenüber Hacking-Angriffen. So ist denkbar, dass durch *Injection*-Attacken sensible Teile der Datenbank, und im schlimmsten Fall sogar die Zugangsverwaltung, ausgelesen werden.[7] Weiterhin stellt der Download größerer Mengen an Dateien im Projektalltag ein gewisses Risiko der ungewollten Weiterverbreitung dar und ist angesichts der notwendigen hohen Dokumentqualität auch recht zeitaufwändig. Eine sinnvolle Alternative ist daher eine Viewer-Anwendung, welche Dokumente bereits serverseitig so gut aufbereitet, dass ein kompletter Download vermieden werden kann. Die Anforderungen solch vergleichsweise komplexer Anwendungen an die Client-Software (i. A. Browser) können jedoch im Laufe der Zeit zu Inkompatibilitäten führen und somit die Nachhaltigkeit der gesamten Anwendung gefährden.

Daher realisiert ArCHO auf der Ebene der Middleware mit Containern und Virtualisierung Architekturprinzipien, wie sie (aus zum Teil sehr verschiedenen Gründen) in der Diskussion um nachhaltige wissenschaftliche Software zur Zeit eine große Rolle spielen. [8] In der konkreten Implementierung wird erreicht, dass der Bildschirm des Nutzers einen per RDP-Protokoll bereitgestellten Virtuellen Desktop zeigt, der jeweils Einzelansichten von Dokumentseiten wiedergibt. [9] Diese können nicht ohne Weiteres heruntergeladen werden. Auch ein programmatischer Zugriff auf die Datenbank ist nicht möglich, daher können Angreifer keinen massenhaften Abzug sensibler Daten erreichen. Außerdem wird die Webanwendung als solche technisch nachhaltig gemacht. Denn da sie sich in einem sehr stark abgeschlossenen System befindet, ist die technische Konfiguration des Client-Rechners zumindest mittelfristig fast ohne Belang.

Die geschilderte Kombination von Virtualisierung, Middleware-Containern und der feingranularen Zugangsverwaltung ist eine pragmatische Antwort auf das ungelöste Problem des rechtssicheren Zugangs zu schutzwürdigen digitalisierten Archivalien. Sie bietet eine Alternative zur räumlichen Zugangsbeschränkung auf Archivlesesäle. ArCHO soll dazu beitragen, die nachhaltige Nutzbarkeit von Daten, die in Forschungsprojekten erhoben wurden, über Orts- und Disziplingrenzen hinweg zu ermöglichen und damit eines der wesentlichen Versprechen der Digitalisierung in den Geisteswissenschaften einzulösen.

Fußnoten

1. Als Beispiel einer über die Geisteswissenschaften hinausgehenden, internationalen Initiative sei die Arbeit der Research Data Alliance genannt: https://rd-alliance.org/ [letzter Zugriff 20. August 2016].
2. S. hierzu die „Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities“, die auch vom MPI für Wissenschaftsgeschichte unterstützt wird: https://openaccess.mpg.de/Berlin-Declaration [letzter Zugriff 20. August 2016].
3. Z. B. das „GESIS Secure Data Center“: http://www.gesis.org/en/services/data-analysis/data-archive-service/secure-data-center-sdc/ [letzter Zugriff 26.11. 2016].
4. ArCHO befindet sich zum Zeitpunkt der Abfassung im Stadium eines Prototypen und wird in der Projektlaufzeit zu einem generischen Service erweitert.
5. http://gmpg.mpiwg-berlin.mpg.de [letzter Zugriff 20. August 2016].
6. Vgl. die Empfehlung 7 der Denkschrift „Sicherung guter wissenschaftlicher Praxis“ der DFG: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf [letzter Zugriff 20. August 2016].
7. Dieses für die meisten historischen Quellenkorpora absurd klingende Szenario ist im Bereich der Zeitgeschichte durchaus als denkbar anzunehmen.
8. Beachte z. B. die thematische Ausrichtung des FORGE-2016-Workshops: https://www.gwiss.uni-hamburg.de/gwin/ueber-uns/forge2016.html [letzter Zugriff 20. August 2016].
9. Die Desktop-Virtualisierung wird mit Apache Guacamole realisiert: https://guacamole.incubator.apache.org [letzter Zugriff 20. August 2016].

Bibliographie

Fornaro , Peter R. / Rosenthaler, Lukas (2016): „File Formats for Archiving: Stability and Persistence Issues“, in: *DH2016: Conference Abstracts* 507–508.

Kruse, Sebastian / Schmaltz, Florian / Stiller, Juliane / Wintergrün, Dirk (2015): „Herausforderung ‚Big Data’ in der historischen Forschung“, in: *DHd 2015: Von Daten zu Erkenntnissen* 171–174 https://dhd2015.uni-graz.at/de/nachlese/book-of-abstracts [letzter Zugriff 20. August 2016].

Neuroth, Heike / Oßwald, Achim / Scheffel, Regine / Strathmann, Stefan / Huth, Karsten (2010): *nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Göttingen: Niedersächsische Staats- und Universitätsbibliothek Göttingen http://www.nestor.sub.uni-goettingen.de/handbuch/index.php [letzter Zugriff 20. August 2016].

Plassmann, Max (2016): „Archiv 3.0? Langfristige Perspektiven digitaler Nutzung“, in: *Archivar. Zeitschrift für Archivwesen* 3: 219–223 http://www.archive.nrw.de/archivar/hefte/2016/Ausgabe_3/Archivar_3_2016.pdf [letzter Zugriff 20. November 2016].

Aufbau eines historisch-literarischen Metaphernkorpus für das Deutsche

Pernes, Stefan
stefan.pernes@uni-wuerzburg.de
Universität Würzburg, Deutschland

Keller, Lennart
jan.keller@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

Peterek, Christoph
christoph.peterek@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

Überblick

Metaphorischer Sprachgebrauch umfasst komplexe gedankliche Würfe genauso wie alltägliche Begrifflichkeiten. Die Metapher gilt als Untersuchungsgegenstand nicht nur in den Literaturwissenschaften, Sprachwissenschaften und der Anthropologie, sondern hat auch Relevanz für so disparate Forschungsprogramme wie das der Künstlichen Intelligenz und der Kritischen Diskursanalyse. Darüber hinaus stellt die Erkennung und Auflösung von Metaphern ein wichtiges Desiderat in sprachtechnologischen Anwendungen dar, deren Gegenstand die Disambiguierung von Wortbedeutungen umfasst. Korpusuntersuchungen zeigen, dass metaphorischer Sprachgebrauch in gängigen Textsorten durchschnittlich in jedem dritten Satz zu finden ist (vgl. Steen et al. 2010; Shutova und Teufel 2010) – ein Beleg für die Ubiquität der Metapher, die in erster Linie darin begründet liegt, dass die idealtypische Karriere eines metaphorischen Ausdrucks als kühne Betonung beginnt und als konventionelle Form endet. Eine Sprachressource zum Metapherngebrauch kann also eine wichtige Ergänzung bei der automatischen inhaltlichen Erschließung von Textbeständen darstellen. Dabei stellt das hier entwickelte Korpus annotierter Sätze, dessen Grundlage eine Sammlung deutschsprachigen Romane aus dem 19. Jahrhunderts bildet, einen spezifischen Beitrag zur Erschließung von historischen Textbeständen dar.

Die große Mehrheit der heute verfügbaren Metaphernkorpora basiert auf dem Prinzip, einige wenige Zielbegriffe sowie unter Umständen ausgewählte konzeptuelle Domänen zu definieren und alle passenden sprachlichen Realisierungen aus einem großen Textbestand zu extrahieren. Diese Herangehensweise lässt vermuten, dass die damit modellierten Eigenschaften sich nicht auf arbiträren Text in realen Anwendungsszenarien übertragen lassen, denn jedes vordefinierte lexikalische oder konzeptuelle Inventar wird dabei zu kurz greifen (vgl. Shutova 2015). Im Gegensatz dazu enthält das hier entwickelte Korpus keine Einschränkungen hinsichtlich der konzeptuellen Domänen oder der erfassten sprachlichen Konstruktionen, bis auf die Tatsache, dass es sich aus literarischen Prosatexten zusammensetzt. Es sollte noch darauf hingewiesen werden, dass mit der Hamburg Metaphor Database eine weitere deutschsprachige Ressource zur Metapher existiert, diese jedoch nach wesentlich anderen Gesichtspunkten erstellt wurde und lediglich eine kleine Zahl ausgewählter Beispielsätze enthält.

Korpuserstellung

Grundlage für die Erstellung des Korpus bildet die Romansammlung der Digitalen Bibliothek des Projektes TextGrid. Die Sammlung umfasst insgesamt 454 Werke vom frühen 16. bis zum frühen 20. Jahrhundert, wobei der Bedarf nach orthographisch normalisiertem Text die Datengrundlage auf 383 Romane aus den Jahren 1830 bis 1940 eingeschränkt hat. Zur Ziehung der zu annotierenden Sätze wird eine balancierte Sampling-Strategie hinsichtlich zeitlicher Streuung und Gender der AutorInnen eingesetzt. Es handelt sich dabei um eine Quotenstichprobe, die aus jedem 10-Jahres-Abschnitt und zu gleichen Teilen männlicher und weiblicher Autorinnen Sätze auswählt. Darüber hinaus wird im Rahmen des Samplings eine automatische Vorauswahl getroffen, sodass die Hälfte der Sätze Metaphern enthält. Möglich wird dies durch einen Classifier, der anhand von TF-IDF Scores – auf Grundlage einer lemmatisieren Version des gesamten Romankorpus – feststellen kann wie “ungewöhnlich” ein zu klassifizierender Satz ist. Anhand eines empirisch festgestellten, von der Größe des TF-IDF Korpus abhängigen, Schwellenwerts ist es anschließend möglich, eine Vorauswahl zu treffen, die indirekt Metaphorizität erfasst. Es handelt sich dabei um eine vereinfachte Form des von Schulder & Hovy

(2014) entwickelten Klassifikationsansatzes. Ziel des hier entwickelten Korpus ist es, einen Gesamtumfang von bis zu 2000 annotierten Sätzen zu erreichen. Als Grundlage dafür wurden insgesamt 3000 Sätze ausgewählt.

Annotation

Wir orientieren uns an der Metaphor Identification Procedure (MIP) der Pragglejaz Group (Pragglejaz Group 2007; Steen et al. 2010) und sehen zunächst jedes Wort im Text als potentielle Metapher. Gegenstand der Metaphernannotation ist es somit, jedes Wort als metaphorisch beziehungsweise nicht metaphorisch zu klassifizieren. Die Aufgabe ist dabei auf metaphorische Äußerungen auf der Wortebene beschränkt, das heißt Satzmetapher und Textmetapher sowie Phänomene grammatischer Metapher sind ausgenommen. Aufgrund der Neigung des Deutschen zur Kompositabildung wird jedoch eine automatische Kompositazerlegung durchgeführt. Da der Umfang der zu annotierenden Sätze eine Herausforderung für eine solche detallierte Herangehensweise wie das MIP darstellt, wird eine automatische Vorselektion potentieller Metaphernkandidaten durchgeführt.

Auf Grundlage von Part-of-Speech-Informationen und Dependency-Bäumen werden aus den Sätzen folgende Konstruktionstypen als Kandidaten für eine metaphorische Verwendung extrahiert (zu den Typen vgl. Skirl und Schwarz-Friesel 2007): Substantivmetapher – dazu gehören Komposita, Kopulakonstruktionen ("X ist ein Y"), Simile ("X ist wie ein Y"), Genitivmetapher – sowie Verb-, Adjektiv-, Präpositions- und 'als'-Metapher. Wir folgen mit diesem Ansatz der automatischen Vorauswahl Gandy et al. (2013), die dadurch bei der Metaphernauszeichnung eine Übereinstimmung der Annotatoren von Kappa = 0.80 erreichen. Die extrahierten Konstruktionen werden anschließend zusammen mit den Sätzen für die Annotation in ein geeignetes Format exportiert und in die Annotationsumgebung WebAnno (Yimam et al. 2014) geladen.

Für den manuellen und bei weitem umfassendsten Teil der Arbeit wurde ein Annotationsleitfaden verfasst, der die Identifikationsstrategie MIP reproduziert, Hinweise zum Umgang mit lexikalisierten metaphorischen Ausdrücken und der Abgrenzung zur Metonymie enthält. Darüber hinaus ist dargestellt, welche Konstruktionstypen vormarkiert werden und welche Ausnahmen dabei zu erwarten sind (Eigennamen und Hilfsverben sind von der Markierung ausgenommen, vgl. Shutova und Teufel 2010). Schließlich wird festgelegt wie mit Ausnahmen und fehlerhaften Sätzen zu verfahren ist. Satzfragmente, starke dialektale Formen sowie Sätze, die ohne Kontext nicht interpretierbar sind, werden als Ausschuss markiert, fehlerhafte Vormarkierungen werden gekennzeichnet und im Rahmen der Kuration der annotierten Sätze verbessert.

Ergebnis

Es kann von den folgenden vorläufigen Ergebnissen der automatischen Vorauswahl und der manuellen Annotation berichtet werden:

Die automatische Extraktion der möglicher Metaphernkandidaten hat es ermöglicht, ein korpusgestütztes Bild darüber zu erlangen wie die Konstruktionen in einer relativ offenen Domäne – einem Romankorpus, das diverse Gattungen umfasst – verteilt sind. Des Weiteren ist ausgehend von der Annotationspraxis festzustellen, dass die erhobenen Konstruktionen – zumindest im Rahmen der hier zugrundeliegenden Texte – theoretisch alle Vorkommen von metaphorischen Äußerungen auf der Wortebene abdecken. In der Praxis kommt es jedoch aufgrund komplexer Hypotaxen und fehlender automatischer Koreferenz-Auflösung zu fehlerhaften Vormarkierungen.

Die vorbereitende, automatische Klassifikation der Sätze in Metapher beziehungsweise Nicht-Metapher führt zu einem Anteil von 48% an Sätzen, die lebendige Metaphern enthalten. Werden lexikalisierte metaphorische Ausdrücke mit eingerechnet, steigt der Anteil der Sätze, die Metaphern enthalten, auf 61%. Ein erheblicher Vorteil, der sich aus der Klassifikation der Sätze ergibt, ist die Fülle des Materials, die sich dadurch generieren lässt. Ohne Vorauswahl liegt die durchschnittliche Anzahl von metaphorischen Ausdrücken pro Satz – je nach Textsorte – zwischen 0.12 und 0.54 (vgl. Shutova & Teufel 2010), während mit dem hier vorgestellten Ansatz ein Wert von 1.91 Metaphern pro Satz erreicht wird. Eine genaue Auswertung der Präzision des Classifiers steht noch aus, in Bezug auf die Struktur der so ausgewählten Sätze kann jedoch festgestellt werden, dass die Klassifizierung keine Auswirkung auf die Verteilung der erhobenen Konstruktionstypen hat.

Für die Übereinstimmung zwischen zwei Annotatoren beim Aufbau des hier vorgestellten

Metaphernkorpus kann ein Wert von 0.87 (Cohen's Kappa) berichtet werden. Werden lediglich die vormarkierten Konstruktionen als Grundlage der Berechnung herangezogen, schwankt Kappa je nach Einbezug lexikalisierter Äußerungen zwischen 0.77 bis 0.80.

Bibliographie

Gandy, Lisa / Allan, Nadji / Atallah, Mark / Frieder, Ophir / Howard, Newton / Kanareykin, Sergey / Koppel, Moshe / Last, Mark / Neuman, Yair / Argamon, Shlomo (2013): „Automatic identification of conceptual metaphors with limited knowledge", in: *Proceedings of AAAI 2013*.

Schulder, Marc / Hovy, Eduard (2014): „Metaphor detection through term relevance", in: *Proceedings of the Second Workshop on Metaphor in NLP*.

Shutova, Ekaterina / Teufel, Simone (2010): „Metaphor corpus annotated for source - target domain mappings", in: *Proceedings of LREC 2010* 3255–3261.

Shutova, Ekaterina (2015): „Design and Evaluation of Metaphor Processing Systems", in: *Computational Linguistics* 41 (4): 579–623.

Skirl, Helge / Schwarz-Friesel, Monika (2007): *Metapher*. Universitätsverlag Winter.

Steen, Gerard J. / Dorst, Aletta G. / Herrmann, J. Berenike / Kaal, Anna A. / Krennmayr, Tina / Pasman, Trijntje (2010): *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam / Philadelphia: John Benjamins.

Yimam, Seid Muhie / Eckart de Castilho, Richard / Gurevych, Iryna / Biemann Chris (2014): „Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno", in: *Proceedings of ACL-2014*, demo session, Baltimore, MD, USA.

Automatische Bild-Text-Analyse: Chancen für die Zeitschriftenforschung jenseits von reinen Textdaten

Rißler-Pipka, Nanette
nanette.rissler-pipka@ku.de
Katholische Universität Eichstätt-Ingolstadt, Deutschland

Chandna, Swati
swati.chandna@kit.edu
Karlsruhe Institute of Technology, Deutschland

Tonne, Danah
danah.tonne@kit.edu
Karlsruhe Institute of Technology, Deutschland

Zeitschriften und multimodale Wahrnehmung

Gerade die Epoche der Moderne (1850-1945) geht mit einer Veränderung der menschlichen Wahrnehmung einher, als mit den aufkommenden Avantgardebewegungen (sowie im Zuge der technischen, ökonomischen und sozialen Weiterentwicklung) weltweit die Anzahl der Kulturzeitschriften explosionsartig steigt. Nicht so sehr die vielzitierte technische Reproduzierbarkeit des Kunstwerks (Benjamin 1935-39) zeugt von dieser Veränderung als vielmehr die Fähigkeit zur multimodalen Wahrnehmung, heraus gebildet durch die dynamische Medienlandschaft zu der neben Photographie und Film vor allem die Zeitschriften gehören. Für Lateinamerika vergleicht Raquel Macciuci die Rezeption der modernen Presse und Zeitschriften mit der Nutzung digitaler Medien (Macciuci 2015: 209), während für den deutschsprachigen Raum gleich zwei Tagungen zu Zeitschriften deren Multimodalität auf der einen Seite („Illustrierte Zeitschriften um 1900: Multimodalität und Metaisierung", 2014) und die Funktionsweise der visuellen Zeitschriftenkultur auf der anderen Seite („Deutsche illustrierte Magazine – Journalismus und visuelle Kultur in der Weimarer Republik", 2013) beleuchten.

Das Desiderat der Forschung wird in dieser Beziehung schon 2010 von Frank et al. genau bestimmt:

> So wissen wir inzwischen relativ gut darüber Bescheid, welche und wie viele einschlägige Zeitschriften es gab und gibt; weniger schon, wer darin worüber aus welcher Perspektive und mit welcher Stoßrichtung geschrieben hat; kaum mehr zuletzt aber, welche Gestaltung die Zeitschriften prägte und welcher Stellenwert dergleichen Publizistik zukam. (Frank et al.: 2010, 10)

Es geht auch hier in der Literatur- und Kulturwissenschaft vor allem um Zahlen, Metadaten und eine Möglichkeit die „Gestaltung" der Zeitschriften quantitativ zu ermessen. Genau diese Felder sind prädestiniert für die Zusammenarbeit mit den DH.

Wenn wir für mehr als 200 Zeitschriftentitel sagen könnten, wie das quantitative Bild-Text-Verhältnis ist und Metadaten dazu vergleichen könnten, wie z.B. die programmatische Ausrichtung des Titels, die Zielgruppe, die Akteure, der Standort und Verbreitungsgrad, dann ließen sich die von Frank et al. gestellten Fragen beantworten. In einer verfeinerten Analyse der Gestaltung kann überprüft werden, ob progressives, avantgardistisches Layout (gekennzeichnet durch viel Leerraum, reduzierte Ornamente, klare Schrifttypen) mit den entsprechend programmatisch eingeordneten Titeln übereinstimmt.

Automatische Bild-Text-Erkennung: ein Versuch mit SWATI

Am konkreten Beispiel mit Daten aus dem Projekt Revistas culturales 2.0 (Universität Augsburg) wurde vor dem Hintergrund dieses Desiderats der Versuch einer automatischen Bild-Text-Erkennung anhand von zunächst 69 Beispielseiten eines Heftes der argentinischen Kulturzeitschrift „El Hogar" (Dec. 1919) unternommen. Während im Projekt die Metadaten der Zeitschriften bereits mit digitalen Tools zur Netzwerkvisualisierung (Ehrlicher / Herzgsell 2016) oder zur Zeit/Ort-Visualisierung mit dem DARIAH-DE Geo-Browser (http://geobrowser.de.dariah.eu/storage/199501) für Überblicksdarstellungen analysiert wurden, fehlte nach wie vor die Möglichkeit einer quantitativen Bild-Text-Analyse, die Gestaltung der Zeitschriften entsprechend.

Der technische Ansatz zur Durchführung einer solchen quantitativen Bild-Text-Analyse basiert auf den Entwicklungen des eCodicology Projektes, in dem die Digitalisate des „Virtuellen Skriptoriums St. Matthias" automatisch ausgewertet und die so erfassten Merkmale in den beschreibenden Metadaten abgelegt und anschließend visualisiert wurden. Auch aus informationstechnologischer Sicht stellt sich die Frage, ob die für mittelalterliche Handschriften verwendeten Verfahren und Algorithmen generisch einsatzbar sind und Potential für alternative Anwendungsfelder, hier am Beispiel spanischsprachiger Magazine untersucht, besitzen. Die Exploration von Erweiterungsmöglichkeiten verspricht ebenfalls zusätzlichen Erkenntnisgewinn für alle beteiligten Disziplinen.

Zur Extraktion der Merkmale wird der SWATI Workflow (Software Workflow for the Automatic Tagging of Medieval Manuscript Images) genutzt (Chandna et al. 2015). Andere bestehende Methoden zur Image Document Analysis führten nicht zum erwarteten Resultat, da sie nur auf spezielle und kleine Datensätze angewendet werden können (vgl. DIVAServices-Spotlight). Hier wurde der SWATI-Workflow speziell an die heterogene Layout-Struktur der Zeitschriften angepasst.

Als Basis der Untersuchung wurden auf den Digitalisaten der spanischsprachigen Magazine die Seitengröße sowie der Text- und Bildraum vermessen und jeweils die Fläche, Breite, Höhe, Koordinaten der linken oberen Ecke sowie der Neigungswinkel bestimmt.

Da auf einer Seite mehrere Text- und/oder Bildbereiche wie beispielsweise Überschriften, Haupttext, Notizen, Initialen, Zeichnungen oder Glossen auftreten können, werden die genannten Werte für jeden Bereich einzeln ermittelt und gespeichert. Zusätzlich werden die Werte auch als relative Angaben in Prozent unabhängig von der Einheit der Messungen aufgeführt, um die Übertragbarkeit bei unterschiedlichen Auflösungen der Digitalisate zu gewährleisten.

Im Bildbeispiel sind die Ergebnisse der jeweiligen automatischen Bild-Text-Erkennung zu erkennen: Originalbild (links) sowie segmentierte Bild- (Mitte) bzw. Textbereiche (rechts). Zusätzlich wird zur Vermessung des Seitenbereiches jeweils eine Segmentierung der Seite durchgeführt (hier nicht dargestellt)


El Hogar

La transparencia del agua

PHILIPS ARGA
la mas economica

HEADINE
El Remedio Más Seguro

Abbildung 1: Originalseite „El Hogar“

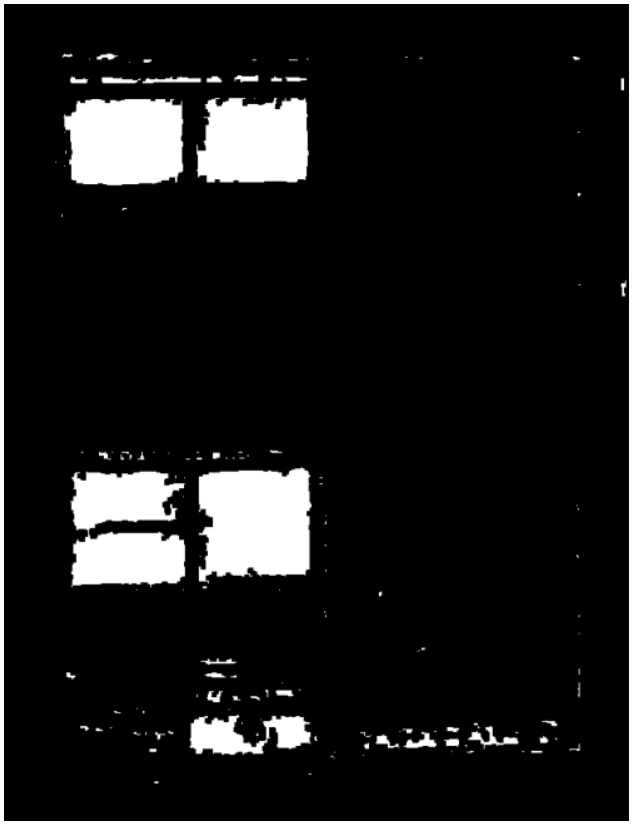


Abbildung 3: Texterkennung

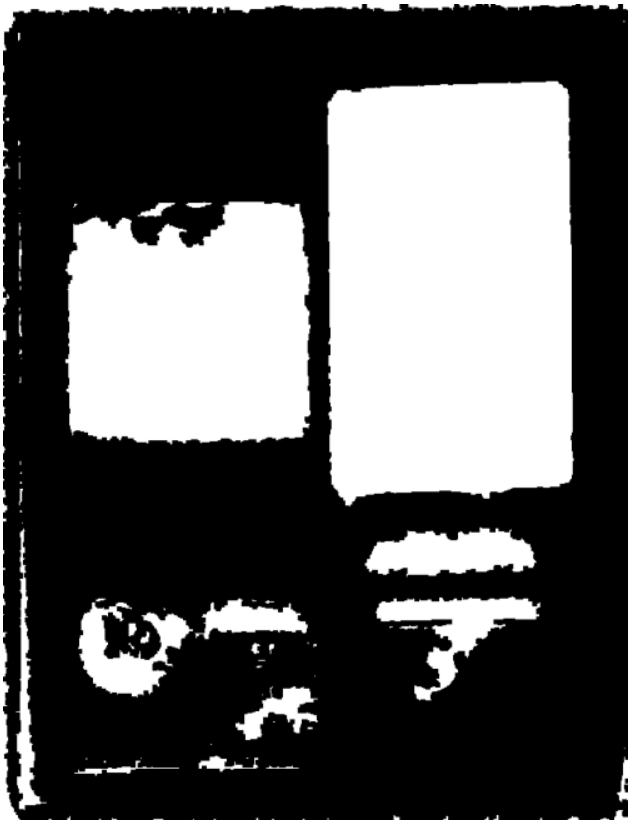


Abbildung 2: Bilderkennung


El Hogar

José Zorrilla

Nuestro Regalo

Para celebrar las Fiestas

Departamento de Lencería

Este es
Nuestro Obsequio

Casa Argentina
Scherrer

Abbildung 4: Originalseite “El Hogar”

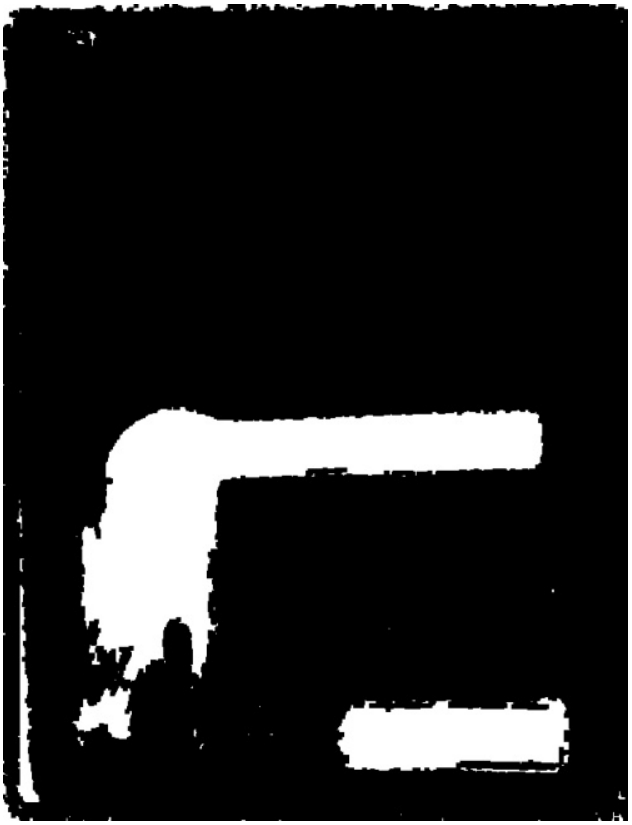


Abbildung 5: Bilderkennung

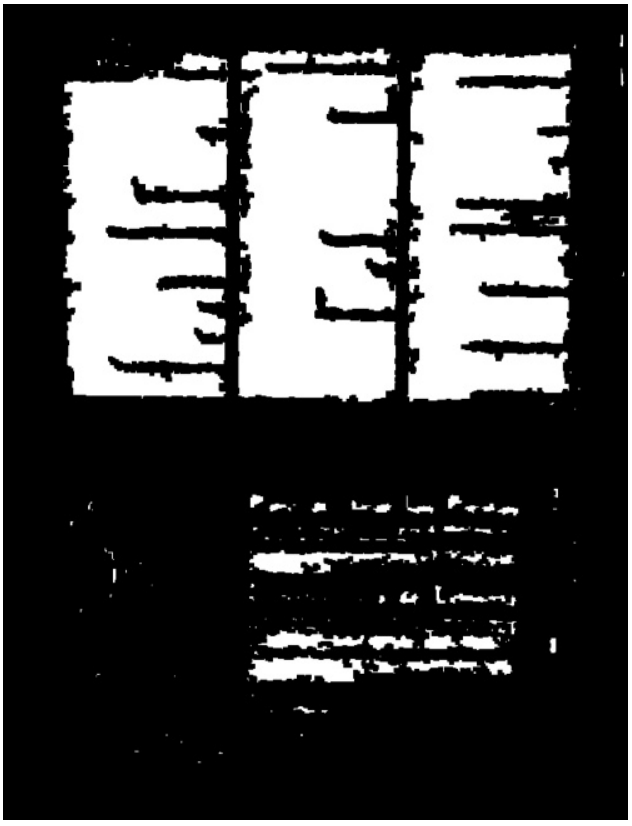


Abbildung 6: Texterkennung

Wie man an diesen beiden Beispielseiten der Zeitschrift „El Hogar“ (Dec. 1919) erkennt, funktioniert die Bild- und Texterkennung sehr zuverlässig auch bei komplexeren Bild-Text-Gefügen, wie z.B. in Werbeanzeigen. Größere ins Bild integrierte Schriften, z.B. die Werbeüberschrift in Abb. 1 oder 4 wird als Teil des Bildes erkannt. Das ist aus gestaltungsanalytischer Sicht auch nicht falsch, hat die Werbeüberschrift doch gleichzeitig Bild- und Textfunktion.

Ein anderes Problem ist die Zuordnung kleinerer Bild- oder Textflächen zu einer Einheit, z.B. gehört in Abb. 4-5 der Schriftblock am unteren rechten Bildrand zum Werbebild und gibt dem zentral stehenden Text eine Rahmung. Als Messergebnis erscheinen aber zwei verschieden große, unabhängige Text- bzw. Bildbereiche.

Die eigentliche Analyse der gewonnenen Metadaten (Messdaten) ist dann für alle Beteiligten eine erneute Herausforderung, kann aber genau das oben beschriebene Problem der Zuordnung semantisch zusammenhängender Text- und Bildbereiche lösen. Durch eine Filtereinstellung werden bei mehr als 10 erkannten Text/Bildbereichen nur die 10 größten ausgewählt. Diese Selektion vermindert den Einfluss von Messartefakten und erleichtert die statistische Auswertung der Daten sowie deren Visualisierung. Auch dieser Bereich wurde im Zusammenhang des eCodicology Projekts bereits für das Korpus mittelalterlicher Handschriften erprobt (Chandna et al. 2016) und konnte auf das vorliegende Fallbeispiel übertragen werden.

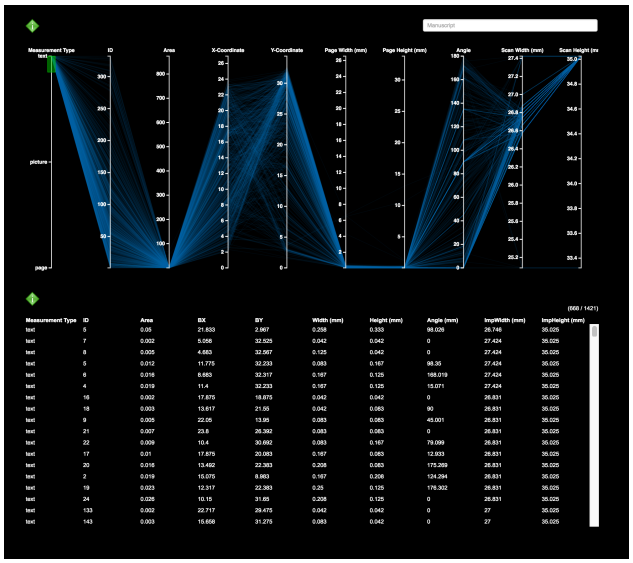


Abbildung 7: Visualisierung der 69 Beispielseiten, hier nur Textbereiche

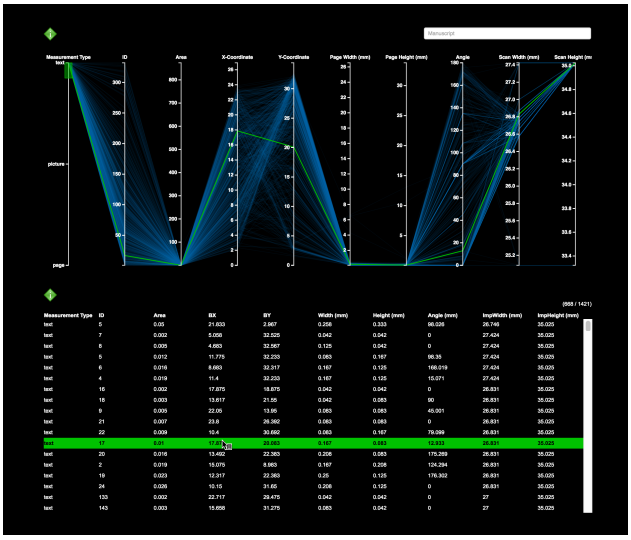


Abbildung 8: Visualisierung der 69 Beispielseiten, nur Textbereiche, Auswahl eines bestimmten Textbereichs (grüne Markierung)

Auf einen Blick sehen wir die Verteilung des Textes, wahlweise im Vergleich mit derjenigen der Bilder auf den Zeitschriftenseiten. Durch die interaktive Funktionalität des Visualisierungstools (CodiVis) können bestimmte Bereiche, einzelne Seiten oder ganze Text- oder Bildblöcke, die bestimmte Gemeinsamkeiten aufweisen, ausgewählt werden.

Wir können auf diese Weise schon beantworten, wie viel Text- und Bildbereiche es im Vergleich auf den gesamten 69 Seiten gibt. Wenn in Zahlen 668 Textbereiche von insgesamt 1421 Bereichen angezeigt werden, ist dabei zwar schon eine Selektion der größeren erfolgt, aber wir können noch nicht sagen, ob 668 einzelne Texte, Spalten, Textteile von Werbeanzeigen oder Überschriften gemeint sind. Umgekehrt sind es 753 Bildbereiche, aber das sagt uns noch nicht, dass es in Summe ebenso viele einzelne Illustrationen oder Werbebilder sind.

Nichtsdestotrotz ist das Ergebnis für einen ersten Versuch mit den Zeitschriftendokumenten erstaunlich. Die automatische Erkennung von Bild und Text funktionierte und es wurden zu jeder einzelnen Seite genaue Metadaten erhoben, die unabhängig von der Annotation einzelner Forscher oder Nutzer sind und insbesondere reproduzierbar erzeugt werden können. Ein solches Verfahren bietet die Möglichkeit, große Korpora automatisiert zu erschließen und so spezifische Fragestellungen an das Material zu ermöglichen.

Fügt man diesen Metadaten die vorhandenen Metadaten (primär Erscheinungszeitraum, Titel, Ort) jeder Zeitschrift hinzu und lässt den beschriebenen Workflow über das gesamte Korpus des Revistas culturales 2.0 Projekts laufen, kann man bereits historische Vergleiche die Gestaltung der Kulturzeitschriften entsprechend anstellen. Konkret erhoffen wir uns, aktuelle Annahmen, wie die Korrelation von bildlicher vs. textueller Gestaltung mit programmatischer Ausrichtung einer Zeitschrift, beantworten zu können. Lassen sich davon Gesetzmäßigkeiten ablesen, z.B. dass avantgardistische Zeitschriften generell mehr nicht-bedruckten Blattanteil haben als andere? Gibt es da Unterschiede, die die Herkunft und das kulturelle Umfeld der Zeitschriften betreffen? Lassen sich bestimmte Muster in der Gestaltung von Zeitschriften erkennen, die regional, personal oder programmatisch zugeordnet werden können? Anhand dieser Fragen können die Parameter, die für die Gestaltungsentscheidung historisch wichtig waren, definiert werden.

Ausblick

Eine automatische Bild-Text-Erkennung und folgende Analyse des gesamten Zeitschriftenkorpus des Revistas culturales 2.0 Projekts ist nach diesem Testlauf der nächste Schritt. Durch Kombination der bereits vorhandenen mit den automatisch erzeugten Metadaten kann die Visualisierungskomponente erweitert und vervollständigt werden. Danach soll geklärt werden, welche Informationen zusätzlich zum reinen Bild-Text-Verhältnis für jede Zeitschrift, Ausgabe, Seite aus den Messdaten gewonnen werden können. Die Position von Bild und Text auf jeder Seite ist neben der Quantität eine ebenso wichtige gestalterische Information. Durch Kontextwissen zu jedem Zeitschriftentitel und den Metadaten können daraufhin Gestaltung und intellektuelles Konzept (also auch Layout und Inhalt) miteinander verglichen werden.

Sowohl aus informationstechnischer als auch aus geisteswissenschaftlicher Sicht ist die Ausweitung der automatisch erkannten Merkmale eine wichtige Herausforderung. Es stellt sich die Frage, welche zusätzlichen Merkmale für eine vollständige Analyse des Bestandes notwendig sind und durch welche angepassten oder zu entwickelnden Algorithmen diese bestimmbar sind. Auch zurzeit fehlerhaft erkannte Elemente könnten durch weiterentwickelte Funktionalitäten verbessert zugeordnet werden.

Langfristig bietet sich eine Analyse aller verfügbaren digitalisierten Zeitschriftentitel (deren Bildqualität ausreicht) an. Somit könnten

nicht nur transatlantisch spanischsprachige Kulturzeitschriften verglichen werden, sondern auch die internationale Szene modernistischer Zeitschriften (vgl. Blue Mountain Project) oder deutschsprachige illustrierte Magazine (vgl. Projekt: “illustrierte Magazine”).

Bibliographie

AsymEnc: http://asymenc.wp.hum.uu.nl [letzter Zugriff 20. November 2016].

Benjamin, Walter (1935-39): „Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit“ in: Tiedemann, Rolf (ed.): *Walter Benjamin. Gesammelte Schriften* 1 (2). Frankfurt am Main: Suhrkamp, 1980, 471–508.

DIVAServices-Spotlight: http://wuersch.pillo-srv.ch/#/ [letzter Zugriff 22. November 2016].

Chandna, Swati / Tonne, Danah / Stotzka, Rainer / Busch, Hannah / Vanscheidt, Philipp / Krause, Celia (2016): „An Effective Visualization Technique for Determining Co-Relations in High-Dimensional Medieval Manuscripts Data“, in: *Visualization and Data Analysis 2016*, San Francisco, California, USA, 14.–18. Februar, 1–6 http://ist.publisher.ingentaconnect.com/contentone/ist/ei/2016/00002016/00000001/art00013

Chandna Swati / Tonne, Danah / Jejkal, Thomas / Stotzka, Rainer / Krause, Celia / Vanscheidt, Philipp / Busch, Hannah / Prabhune, Ajinkya (2015): „Software workflow for the automatic tagging of medieval manuscript images (SWATI)“, in: Ringger, Eric K. / Lamiroy, Bart (eds.): *Proceedings SPIE9492, Document Recognition and Retrieval* XXII, 940201 (8. Februar 2015), San Francisco.

Chinese Women's Magazines: http://kjc-sv013.kjc.uni-heidelberg.de/frauenzeitschriften/index.php [letzter Zugriff 20. November 2016].

Chinesische Unterhaltungspresse: http://projects.zo.uni-heidelberg.de/xiaobao/index.php?p=start [letzter Zugriff 20. November 2016].

Die Fackel, Austrian Academy Corpus: http://corpus1.aac.ac.at/fackel [letzter Zugriff 20. November 2016].

eCodicology Projekt: http://www.ecodicology.org [letzter Zugriff 20. November 2016].

Ehrlicher, Hanno / Herzgsell, Teresa (2016): „Zeitschriften Als Netzwerke Und Ihre Digitale Visualisierung: Grundlegende Methodologische Überlegungen Und Erste Anwendungsbeispiele“, in: *Revistas Culturales 2.0*. http://www.revistas-culturales.de/de/buchseite/hanno-ehrlicher-teresa-herzgsell-zeitschriften-als-netzwerke-und-ihre-digitale [letzter Zugriff 20. November 2016].

ESPrit: http://www.espr-it.eu [letzter Zugriff 20. November 2016].

Europeana Newspapers: http://www.europeana-newspapers.eu [letzter Zugriff 20. November 2016].

illustrierte magazine: http://magazine.illustrierte-presse.de [letzter Zugriff 20. November 2016].

Frank, Gustav / Podewski, Madleen / Scherer, Stefan (2010): „Kultur – Zeit – Schrift. Literatur- und Kulturzeitschriften als ‚kleine Archive‘“, in: *Internationales Archiv für Sozialgeschichte der deutschen Literatur (IASL)* 34 (2): 1–45.

Macciuci, Raquel (2015): „Técnica, soporte, ámbitos de sociabilidad y mecanismos de legitimación: sobre la construcción de espacios de literatura en la prensa periódica“, in: Schlünder, Susanne / Macciuci, Raquel (eds.): *Literatura y técnica: derivas ficcionales y materiales: Libros, escritores, textos, frente a la máquina y la ciencia.* Actas del VIII Congreso Orbis Tertius. La Plata: Ediciones del lado de acá 205–231.

Revistas culturales 2.0: Virtuelle Forschungsumgebung zur Erforschung spanischsprachiger Kulturzeitschriften der Moderne (2014–2016): Universität Augsburg. 2014-2016. http://www.revistas-culturales.de [letzter Zugriff 20. November 2016].

Rißler-Pipka, Nanette (2014): „Sobre Los Problemas de Investigación Con Revistas Culturales Digitalizadas Del Mundo Hispanohablante“, in: Rißler-Pipka, Nanette / Ehrlicher, Hanno (eds.): *Almacenes de un tiempo en fuga: Revistas culturales en la modernidad hispánica*. Aachen: Shaker 59–80.

Virtuelles Skriptoriums St. Matthias: http://stmatthias.uni-trier.de [letzter Zugriff 20. November 2016].

WeChangEd: http://www.wechanged.ugent.be [letzter Zugriff 20. November 2016].

Yang, Tze-I / Torget, Andrew / Mihalcea, Rada (2011): „Topic Modeling on Historical Newspapers“, in: *Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities LaTeCH*, 24. Juni 2011 Portland, Oregon, USA, Proceedings of the Workshop, Association for Computational Linguistics, Stroudsburg, USA, 96–104. https://www.aclweb.org/anthology/W/W11/W11-15.pdf [letzter Zugriff 20. November 2016].

ZEFYS: http://zefys.staatsbibliothek-berlin.de [letzter Zugriff 20. November 2016].

Autorschaftsattribution bei nicht-normalisiertem Mittelhochdeutsch. Bessere Erkennungsquoten durch ein Normalisierungswörterbuch

Dimpel, Friedrich Michael
mail@dimpel.de
FAU Erlangen-Nürnberg, Deutschland

Einleitung: Delta im Mittelalter und in der Forschung

Im Bereich der Autorschaftsattribution sind in den letzten Jahren große Fortschritte erzielt worden; insbesondere der Delta-Test nach Burrows' (2002) und Varianten zu Burrows' Verfahren haben sich in vielen Validierungsstudien als sehr erfolgreich erwiesen (Hoover 2004, Eder / Rybicki 2011, Eder 2013a, Eder 2013b, Jannidis / Lauer 2014, Evert / Proisl / Jannidis / Pielström / Schöch / Vitt 2015, Evert / Proisl / Jannidis / Pielström / Reger/ Schöch / Vitt 2016). In mittelalterlichen Texten stellen sich jedoch besondere Probleme: Hier ist die Schreibung weitgehend nicht normiert, das Wort ‚und' wird teilweise mit ‚u' oder ‚v', mit weichem ‚d' oder hartem ‚t' geschrieben; die Genitiv-Form zum nhd. Wort ‚Gott' lautet ‚gotes' oder ‚gotis' (Viehhauser 2015).

Im Rahmen eines Vortrags auf der DHd-Tagung 2016 in Leipzig konnte ich zeigen (Dimpel 2016), dass Delta bei normalisierten mittelhochdeutschen Texten sehr gut funktioniert, insbesondere dann, wenn Texte verwendet werden, die aus mindestens 5.000 Wörtern bestehen, und wenn die Bag-of-Words-Technik (vgl. Eder 2013b) zum Einsatz kommt. Um die Erkennungsquote zu ermitteln, habe ich ein „Ratekorpus" und ein „Validierungskorpus" gebildet. In beiden Sammlungen sind Texte mit bekannter Autorschaft enthalten. Zu jedem Autor, der im Ratekorpus enthalten ist, ist jeweils ein Text des gleichen Autors im Validierungskorpus enthalten. Ermittelt wurde der Prozentsatz der richtig erkannten Autoren. Bei einem Test mit 16 Texten im Validierungskorpus und 15 Texten im Ratekorpus wurde eine Erkennungsquote von 97,1% ermittelt.

Erster Validierungstest nicht-normalisierte Texte

Zu nicht-normalisierten Texten habe ich 2016 erste Zahlen ebenfalls mit positivem Ergebnis vorlegen können, die allerdings nicht valide sind, weil mir zu diesem Zeitpunkt nur sehr wenige nicht-normalisierte Texte digital verfügbar waren: bei einer Textlänge von 5.000 Wörtern konnte ich gegen ein Validierungskorpus mit 14 Texten nur 6 Texte von 5 Autoren prüfen. Nunmehr liegen weitere Texte vor, so dass für einen Validierungstest nun ein Ratekorpus mit 15 Texten von 10 Autoren zur Verfügung steht. Im Validierungskorpus ist je ein Text dieser 10 Autoren enthalten, dazu kommen weitere 10 Texte, die Fehlattributionen auslösen könnten.

Dass Delta bei nicht-normalisierten Texten schlechtere Erkennungsquoten liefert, ist deshalb zu erwarten, weil Delta auf der Verteilung von hochfrequenten Wörtern beruht. Wenn im Werk X überwiegend ‚unt' steht, wenn sich im Werk Y des gleichen Autors jedoch der Abschreiber für die Graphie ‚vnnd' entschieden hat, wird die Zuordnung des richtigen Autors dadurch erschwert. Erwartungsgemäß liegt die Erkennungsquote mit ca. 80% (bei Bag-of-Words mit 5.000 Wörtern; 50 Iterationen zum Ausgleich von Zufallsschwankungen bei der Bag-of-Words-Bildung; davon der Mittelwert für die Vektoren 200, 400, 600 und 800) deutlich unter der Quote für normalisierte Texte. Um eine Verbesserung der Erkennungsquote zu ermöglich, wurden nun Ansätze zur automatischen Teilnormalisierung erprobt.

Teilnormalisierung: Normalisierungswörterbuch und Vollformenwörterbuch

Ein erster Schritt dabei ist die Eliminierung der Sonderzeichen und der deutschen Umlaute – auch in dem soeben erwähnten Test war diese Bereinigung bereits implementiert. Eine weitere automatische Teilnormalisierung ist deshalb ökonomisch realisierbar, weil für den Delta-Test keine vollständige Normalisierung nötig ist. Weil

Delta auf den hochfrequenten Wörtern beruht, sollte bereits eine Normalisierung der häufigen Wörter zu einer Verbesserung führen.

In einem nächsten Schritt wurde ein Skript entwickelt, das aus einigen kürzeren nicht-normalisierten Texten die hochfrequenten Wortformen heraussucht und den User bittet, die normalisierte Form zuzuordnen. Eine Vorschlagsliste aus einem normalisierten Lachmann-Korpus, die mittels Levenshtein-Distanz generiert wurde, hat meiner Hilfskraft das Leben leichter gemacht. Zudem habe ich zwei Datengeschenke bekommen: Sonja Glauch hat mir Daten aus dem Projekt „Lyrik des Mittelalters“ gegeben, das eine Zuordnung von normalisierten zu nicht-normalisierten Wortformen herstellt. Mit den Skript-Daten und den Lyrik-Projekt-Daten lag ein Normalisierungswörterbuch mit gut 1.100 Zuordnungen vor, als mir Thomas Klein Daten aus dem Referenzkorpus Mittelhochdeutsch überlassen hat. Die vorbildliche Struktur des Referenzkorpus hat es möglich gemacht, weitere Zuordnungen zu extrahieren und sie in das Normalisierungswörterbuch einspeisen, das nunmehr gut 120.000 Zuordnungen enthält.

Eine Sichtung des Normalisierungswörterbuchs hat jedoch gezeigt, dass teilweise auch solche diplomatische Wortformen wie ‚sluc‘ zu ‚sluoc‘ normalisiert werden, die eigentlich auch selbst als normalisierte Form eines anderen Lemmas stehen könnten: ‚sluc‘ kann als starkes Femininum etwa nhd. „ein Schluck“ heißen und müsste dann nicht durch eine andere normalisierte Form ersetzt werden. Mitunter wurde im Lyrikprojekt und im ReM unerwartet normalisiert: So findet sich bspw. eine Normalisierung der Wortform ‚chunich‘ zu ‚küninc‘, während im BMZ und im Lachmann-Parzival meist ‚künec‘ steht.

Um das Normalisierungswörterbuch zu überprüfen und vereinheitlichen zu können, wurde ein mittelhochdeutsches Vollformenwörterbuch benötigt, das die Wortformen enthält, die zu normalisierten Lemmata durch Flexion gebildet werden können.

Aus der CD „Mittelhochdeutsche Wörterbücher im Verbund“ (Trier, Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften 2002) wurden Daten extrahiert und mögliche Flexionsformen zu den Lemmata generiert. Wenn es auf Vollständigkeit und Korrektheit ankommen würde, wäre die Erstellung eines derartigen Vollformenwörterbuchs eine große Herausforderung. Doch geht es sowohl bei dem Normalisierungswörterbuch als auch bei dem Vollformenwörterbuch hier nur darum, eine prozentuale Verbesserung der Delta-Erkennungsquote zu erreichen. Fehler bei seltenen Wortformen sind bei Delta meist zu vernachlässigen, wichtig ist eine Vereinheitlichung der häufigen Wörter. Manche Probleme haben sich überwiegend erfreulich lösen lassen: Zu Ablaut und grammatischem Wechsel sind Informationen BMZ hinterlegt. Schwache Verben mit sogenanntem Rückumlaut generiert das Skript dann, wenn im Singular Präsens ein umgelauteter Vokal sowie Positionslänge oder Naturlänge vorliegen. In der verfügbaren Zeit wurde Vieles nicht vollständig gelöst – für Nomina mit Umlaut müssten im Artikel noch die Belegstellen examiniert werden, hier wird bislang nur der Artikelkopf des Lexer ausgewertet. Funktionswörter sind überwiegend listenbasiert ergänzt.

Eine Evaluierung hat gezeigt, dass eine vollständige Bereinigung des Normalisierungswörterbuchs um Formen wie ‚sluc‘ zu einer minimalen Verschlechterung der Erkennungsquote führt, so dass die Eliminierung von diplomatischen Wortformen, die auch normalisierte Wortformen sein könnten, bei hoch- und mittelfrequenten Wortformen nicht angewendet wurde.

Delta-Verbesserung: Z-Wert-Begrenzung

Bei Delta berechnet man bspw. für 200 Most-Frequent-Words für jedes dieser Worte einzeln Z-Werte, in die die Abweichung der Häufigkeit eines Wortes in einem Text zur Häufigkeit dieses Wortes im Gesamtkorpus unter Berücksichtigung der Standardabweichung eingeht. Delta ist das arithmetische Mittel der positiven Z-Wert-Differenzen (Burrows 2002). Evert / Proisl / Jannidis / Pielström / Reger/ Schöch / Vitt 2016 haben meinen Verdacht evaluiert, dass Delta weniger aufgrund einzelner Extremwerte funktioniert (Ausreißerhypothese), sondern eher aufgrund einer breiten autorspezifischen Verteilung der Z-Werte (Schlüsselprofilhypothese).

Fehlt eine Wortform in einem Text, kann dies mitunter mit erhöhten negativen Z-Werten einhergehen. Evert et alia haben besonders hohe Z-Werte auf einen Maximalwert begrenzt, so dass „Ausreißer“ nur abgemildert eingehen. Wenn der Erfolg von Delta den „Ausreißern“ zu verdanken wäre, hätte sich die

Erkennungsquote bei einer Begrenzung der Z-Werte verschlechtern müssen. Ein Begrenzen der Z-Werte führt jedoch zu einer Verbesserung der Erkennungsquote. Evert et alia haben so nicht nur die Schlüsselprofilhypothese bestätigen können, sondern zugleich eine Möglichkeit entdeckt, die Erkennungsquote zu verbessern, die gerade bei mittelalterlichen Texten nützlich sein kann: Wenn in einem Text ein Schreiber eine bestimmte Schreibvariante ganz vermeidet, können Nullwerte zu hohen Z-Werten führen; diese Schreibvariantenproblematik kann durch das Begrenzen der Z-Werte gemildert werden.

Zweiter Validierungstest nicht-normalisierte Texte

Bei dem zweiten Validierungstest habe ich einerseits nicht-normalisierte Wortformen mit Hilfe des Normalisierungswörterbuchs bei der Erstellung der jeweiligen Bag-of-Words in eine normalisierte Wortform konvertiert. Andererseits habe ich eine Z-Wert-Begrenzung durchgeführt und Z-Werte ab |1,64| auf den Wert 1,70 gesetzt (dieser Wert hat sich in einer Versuchsreihe mit verschieden gestalteten Validierungs- und Ratekorpora als vorteilhaft erwiesen). Die Erkennungsquote für Bag-of-Words mit 5.000 Wortformen steigt damit von 80% auf 91% an.

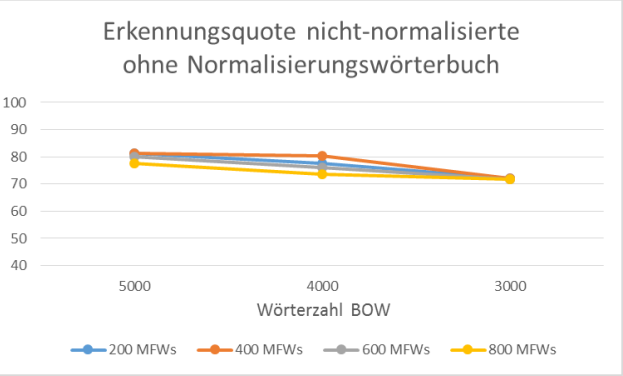


Abb. 1: Erkennungsquoten ohne Normalisierungswörterbuch / Z-Wertbegrenzung für nicht-normalisierte Texte (15 Texte Ratekorpus / 20 Texte Validierungskorpus)

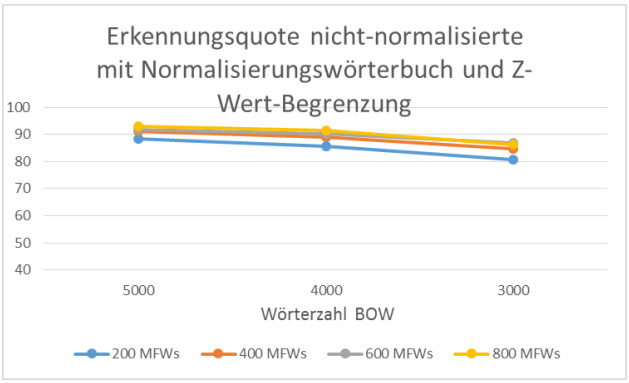


Abb. 2: Erkennungsquoten mit Normalisierungswörterbuch / Z-Wertbegrenzung für nicht-normalisierte Texte (15 Texte Ratekorpus / 20 Texte Validierungskorpus)

Wenn man die mühevoll bereinigten Texte nun wieder mit Fehlern kontaminiert, indem man den Inhalt der Bag-of-Words bspw. durch korpusfremdes Vokabular (teilweise durch altfranzösische Wörter statt mhd. Wörter) austauscht, so sinkt die Erkennungsquote erstaunlich langsam. Wenn 12% der Wörter durch Fremdmaterial getauscht wurden, ist nur ein geringes Absinken erkennbar. Tauscht man 20% aller Wörter durch Noise aus, dann gibt die Erkennungsquote etwas mehr nach als bei normalisierten Texten (vgl. Dimpel 2016) – das ist plausibel, weil hier trotz aller Anstrengungen mit Normalisierungswörterbuch und Z-Wert-Begrenzung noch immer mehr Varianz in den Texten enthalten ist als in Texten, die ein Editor manuell normalisiert hat. Dennoch bleiben die Quoten erstaunlich stabil.

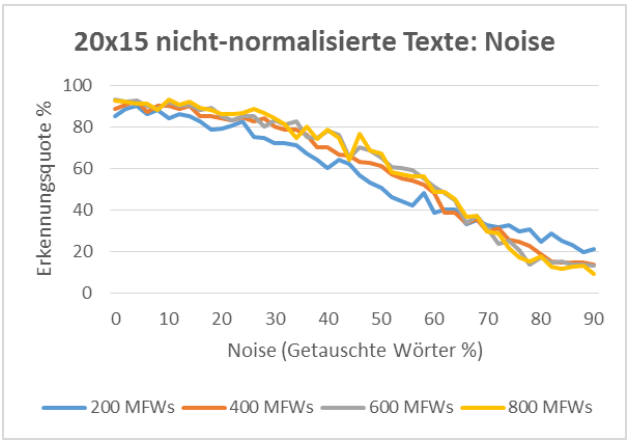


Abb. 3: Noise – Absinken der Erkennungsquote beim Tausch des Wortmaterials der BOW der Ratedatei

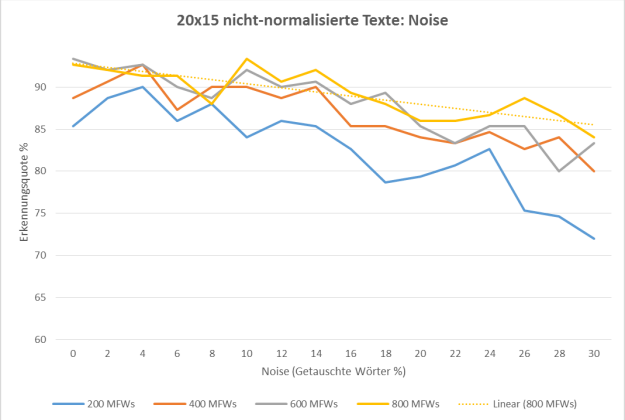


Abb. 4: Noise – Ausschnittvergrößerung 0-30 % Noise

Bibliographie

Burrows, John (2002): „‚Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship", in: *Literary and Linguistic Computing* 17 (3): 267–87 10.1093/llc/17.3.267.

Dimpel, Friedrich Michael (2016): „Burrows' Delta im Mittelalter: Wilde Graphien und metrische Analysedaten", in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 65–70.

Eder, Maciej (2013a): „Mind Your Corpus: systematic errors in authorship attribution", in: *Literary and Linguistic Computing* 28: 603–614 10.1093/llc/fqt039.

Eder, Maciej (2013b): „Does size matter? Authorship attribution, small samples, big problem", in: *Literary and Linguistic Computing Advanced Access* 29: 1–16 10.1093/llc/fqt066.

Eder, Maciej / Rybicki, Jan (2011): „Deeper Delta across genres and languages: do we really need the most frequent words?", in: *Literary and Linguistic Computing* 26 (3): 315–321 10.1093/llc/fqr031 .

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): „Towards a better understanding of Burrows's Delta in literary authorship attribution", in: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, CO: Association for Computational Linguistics, 79–88 10.5281/zenodo.18177 http://www.aclweb.org/anthology/W/W15/W15-0709.pdf [letzter Zugriff 20. August 2015].

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Pielström, Steffen / Reger, Isabella/ Schöch, Christof / Vitt, Thorsten (2016): „Burrows Delta verstehen", in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 61–65.

Hoover, David L. (2004): „Delta Prime?", in: *Literary and Linguistic Computing* 19 (4): 477–495 10.1093/llc/19.4.477 .

Jannidis, Fotis / Lauer, Gerhard (2014). „Burrows's Delta and Its Use in German Literary History", in: Erlin, Matt / Tatlock, Lynne (eds.): *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*. New York: 29–54.

Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): „Improving Burrows' Delta - An Empirical Evaluation of Text Distance Measures", in: *DH2015: Global Digital Humanities* http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empi/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empirical_

Viehhauser, Gabriel (2015): „Historische Stilometrie? Methodische Vorschläge für eine Annäherung textanalytischer Zugänge an die mediävistische Textualitätsdebatte", in: Baum, Constanze / Stäcker, Thomas (eds.): *Grenzen und Möglichkeiten der Digital Humanities*. Sonderband ZfdG 1.

Bild, Beschreibung, (Meta)Text Automatische inhaltliche Erschließung und Annotation kunsthistorischer Daten

Dieckmann, Lisa
lisa.dieckmann@uni-koeln.de
Universität zu Köln, Deutschland

Hermes, Jürgen
hermesj@uni-koeln.de
Universität zu Köln, Deutschland

Neuefeind, Claes
c.neuefeind@uni-koeln.de
Universität zu Köln, Deutschland

Der Vortrag thematisiert die automatische inhaltliche Erschließung und linguistische Annotation der digitalen Repräsentationen kunst- und kulturhistorischer Artefakte innerhalb des prometheus Bildarchivs (http://prometheus-bildarchiv.de), das derzeit 89 Datenbanken aus Museen und Forschungsinstitutionen mit insgesamt über 1,5 Mio. Datensätzen zusammenführt

(Dieckmann 2015). Das durch eine bereits abgeschlossene Vorstudie initiierte Projekt verfolgt zwei unterschiedliche, teilweise aufeinander aufbauende Ansätze: Zum einen die Annotation von Freitexten zur strukturierten Erschließung kunsthistorischer Daten, zum anderen die Analyse der Identität von Datensätzen über die Berechnung gradueller Ähnlichkeiten von Objekten. Beide Ansätze dienen erstens einer Verbesserung des Retrievals, zweitens einer nachhaltigen Sicherung der Daten durch die Verknüpfung mit Normdaten; drittens sollen die zusätzlich erschlossenen Informationen längerfristig als Grundlage für weiterführende (fachspezifische) Fragestellungen eingesetzt werden, etwa zur Rekonstruktion von Künstlergruppen durch die Erstellung von Personen-Netzwerken. Das Projekt wird an der Universität zu Köln in enger Zusammenarbeit zwischen Fachwissenschaftlern der Kunstgeschichte und der Sprachlichen Informationsverarbeitung (http://www.spinfo.phil-fak.uni-koeln.de/) durchgeführt, deren Schwerpunkte u.a. auf Systemen zur syntaktischen und semantischen Analyse und Verarbeitung textueller Daten (Hermes 2012, Schwiebert 2012) sowie zur Annotation nicht-standardisierter Daten (Neuefeind 2013) liegen.

Metadaten und Referenzobjekte

Die digitalen Repräsentationen der in prometheus zusammengeführten kunst- und kulturhistorischen Artefakte stellen insofern eine besondere Herausforderung für die automatisierte Erschließung inhaltlicher Informationen dar, als dass die Metadaten und Texte strukturell und inhaltlich sehr heterogen und in unterschiedlichen Kontexten vorliegen. Die Datensätze der einzelnen Bilddatenbanken sind zwar stets in ein eigenes Metadatenschema eingepasst, jedoch erfolgt die Erschließung der Werke an den jeweiligen Institutionen nicht nach einer einheitlichen Methodik, was u.a. datenbank- oder sammlungsspezifische Gründe hat. Zum einen liegt innerhalb der Klassifikationen der jeweiligen Datenbanken eine Vielzahl an Texten vor, die bislang nicht strukturiert erschlossen sind, sondern derzeit nur über eine einfache Volltextsuche miteinbezogen werden (es handelt sich hierbei oftmals um unstrukturierte Freitextfelder, die z.B. Angaben über Standort(e), die Publikationsgeschichte oder ausführliche Bildbeschreibungen enthalten können). Zum anderen wird selten ein bestimmter Metadatenstandard zugrunde gelegt oder auf Fachvokabulare und Terminologieressourcen zurückgegriffen, was dazu führt, dass zum Teil stark variierende Schreibweisen u.a. bei Künstler- oder Ortsnamen existieren. In der kunsthistorischen Forschung haben sich zudem selten einheitliche Bezeichnungen für Werktitel durchgesetzt. So liegt bspw. das Werk “Bonaparte überquert den großen Sankt Bernhard” von Jacques-Louis David (Malmaison, 1801) in prometheus in mindestens sieben verschiedenen Titelbezeichnungen vor, die zumeist in Teilen, unter Umständen aber auch vollständig voneinander abweichen (etwa “Napoleon überquert die Alpen” gegenüber “Bonaparte auf dem großen Sankt Bernhard”). Eine Verknüpfung mit Normdaten wie der Gemeinsamen Normdatei der Deutschen Nationalbibliothek (GND, http://d-nb.info/gnd/1067141367) ist auf dieser Grundlage nicht möglich. Diese wäre aber nötig, um eine automatische Zusammenführung der Einzelabbildungen zu Objekten vornehmen und die Objekte eindeutig und damit nachhaltig identifizieren zu können, was zugleich die Grundlage für eine weitere Anreicherung mit GND-verknüpften Daten oder weiteren Normdaten (z.B. VIAF, http://viaf.org ; Wikidata, https://www.wikidata.org) bilden würde.

Methodologie

Die Heterogenität der Daten wird in prometheus bereits teilweise in Anwendung linguistischer Analyseverfahren bei der Indexierung ausgeglichen, wobei der Schwerpunkt hier v.a. auf der orthographischen und morphosyntaktischen Ebene liegt, etwa auf Grundlage sprachspezifischer Wörterbücher (u.a. zur Grundformreduzierung, Phrasenerkennung, Synonymgenerierung, Kompositazerlegung) sowie durch Anreicherung mit synonymen Künstlernamen (siehe http://prometheus-bildarchiv.de/tools/pkn d). Diese Maßnahmen dienen in erster Linie dazu, das Retrieval zu optimieren und den Recall zu verbessern. In Bezug auf die oben aufgeworfenen Probleme der Normalisierung und Zuordnung von Einzeldarstellungen zu Objekten sind sie jedoch nur als ein erster Schritt anzusehen. Ziel ist vielmehr ein erweiterter Thesaurus, in dem die tatsächlich auftretenden, zum Teil stark variierenden Schreibweisen von Werktiteln und Künstlernamen auf die verfügbaren Normdaten abgebildet werden.

Da die Variation in den Schreibweisen keine eindeutige Zuordnung erlaubt, bedarf es hierbei zusätzlicher Kriterien. Im Zuge des Projekts wird hierfür ein semantisch motiviertes Verfahren erarbeitet, das die gesamten zu einem Objekt verfügbaren Informationen berücksichtigt: Neben den bereits erschlossenen Metainformationen (wie Name, Titel, Datierung, Standort, etc.) sollen auch die in den bislang nur unstrukturiert vorliegenden Freitextfeldern (s.o.) enthaltenen Informationen genutzt werden können. Zu diesem Zweck werden die Texte zunächst mittels Informationsextraktion aufbereitet (Annotation von Orts- und Personennamen, Zeitausdrücken, etc). Auf Grundlage dieser neu gewonnenen Informationen werden zusätzliche, das Objekt beschreibende Merkmale erstellt und in Form von Feature-Vektoren kodiert (Features sind z.B. „Personen“, „Orte“, „Material“, o.ä.; Werte sind jeweils die konkreten Nennungen, vgl. Abb. 1).

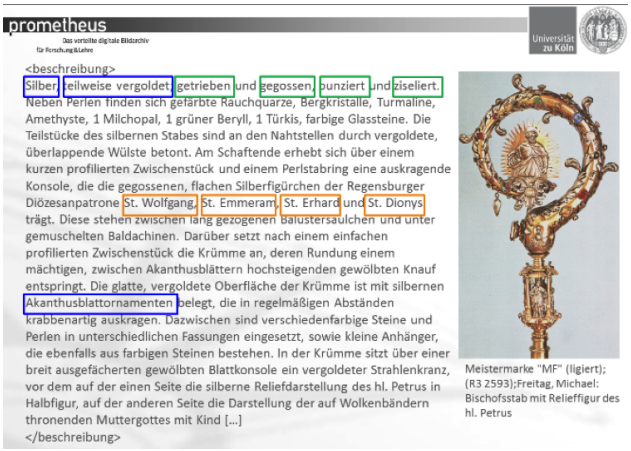


Abb. 1: Beispiel einer Freitextbeschreibung im prometheus-Bildarchiv, in der exemplarisch mittels Informationsextraktion identifizierte Elemente markiert wurden.

Aus den zusätzlichen Merkmalen kann nun, in Kombination mit den bereits vorhandenen Metainformationen, für jedes Objekt ein „semantisches Profil“ bzw. „Fingerprint“ erstellt werden, anhand dessen sich die Ähnlichkeit zwischen Objekten ermitteln lässt. Die Ähnlichkeit wird dabei zunächst in Bezug auf die einzelnen Merkmale ermittelt (u.a. mittels Edit-Distance oder Soundex- bzw. Metaphone-Difference zwischen einzelnen Feldern, Abgleich zeitlicher Angaben, Distanz zwischen Feature-Vektoren zu „Personen“, „Orten“, „Material“, etc.), wobei der Einfluss einzelner Merkmale unterschiedlich gewichtet werden kann. Daraus wird ein kombiniertes Maß der Übereinstimmung zwischen zwei Datensätzen errechnet, das auch bei deutlich abweichenden Schreibweisen eine Aussage darüber erlaubt, ob es sich um das gleiche Objekt handelt. Auf dieser Grundlage können identische Objekte dann auf das jeweilige Referenzobjekt der GND abgebildet werden.

In einem vorbereitenden Projekt für das laufende Vorhaben wurden zunächst die bestehenden Metadaten der einzelnen Datenbanken des prometheus-Bildarchivs quantitativ ausgewertet, um einen Überblick darüber zu erlangen, wie sich der Umfang der zu erschließenden Daten darstellt. Die meisten der 89 Datenbanken verfügen über noch nicht erschlossene Freitextbeschreibungen der Objekte. Diese erstrecken sich zu einem nicht geringen Teil über mittellange (25-75 Wörter) und lange (>75 Wörter) Texte, die im Zuge des Projekts aufbereitet werden sollen. Abb. 2 zeigt die Verteilung dieser unterschiedlichen Textsorten in ausgewählten Datenbanken. Einige verfügen über keinerlei Freitext-Bildbeschreibungen, z.B. die Datenbank des *Zentralarchivs für Kunstgeschichte* in München (zi_muc). Andere, etwa die Erlanger Datenbank *Zeichnungen der graphischen Sammlung* (erlangen_z), weisen fast ausschließlich kurze Beschreibungen auf, wieder andere enthalten dagegen auch eine Reihe mittellanger und langer Texte.

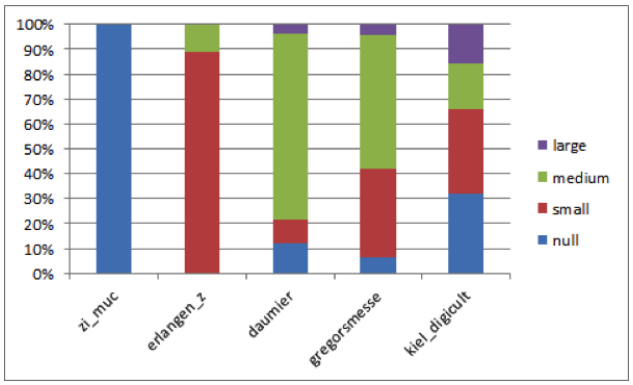


Abb. 2: Verteilung der Freitextlängen über verschiedene Datenbanken des prometheus-Bildarchivs

Zur Nutzung der in den Bildbeschreibungen und Ikonographien enthaltenen Informationen müssen diese zunächst identifiziert und entsprechend ausgezeichnet werden. Dafür wurde zunächst ein Komponenten-Workflow konzipiert und auf Basis des UIMA-Frameworks (Unstructured Information Management Architecture, siehe https://uima.apache.org) implementiert. Im Zuge der Verarbeitung werden die zu annotierenden Informationen

in ausgewählten Feldern der Datensätze identifiziert (vgl. Abb. 3). Dabei kommen Standardmethoden der Informationsextraktion (z.B. Temporal Expression Detection, Named Entity Recognition) genauso zum Einsatz wie informationstheoretische Maße (etwa Log-Likelihood oder tf.idf), um domänenspezifisch relevante Terme zu bestimmen.

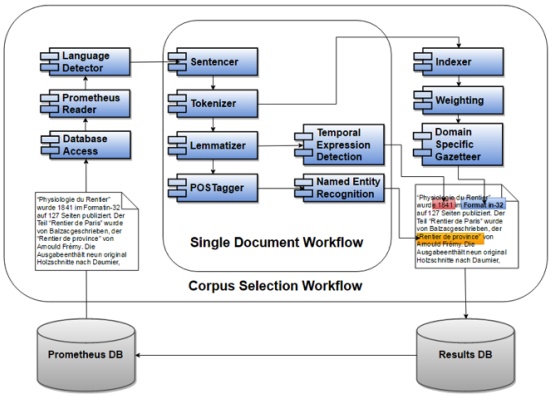


Abb. 3: Workflow zur Informationsextraktion in Freitexten im prometheus-Bildarchiv.

Die erste Projektphase diente vor allem der Evaluierung von Werkzeugen, etwa dem Stanford Named Entity Tagger (siehe http://nlp.stanford.edu/software/CRF-NER.shtml) zur Identifikation von (Orts-)Namen, oder HeidelTime (siehe http://dbs.ifi.uni-heidelberg.de/index.php?id=129) zur Annotation von Zeitausdrücken, um den voraussichtlichen Bedarf an Anpassungen der Werkzeuge für die kunsthistorische Domäne zu ermitteln. Abb. 4 zeigt das Ergebnis der StanfordNER-Komponente, die in einem Text der Datenbank “The Daumier Register” (http://www.daumier-register.org/) Eigennamen auszeichnet: Künstlernamen (“Henry Monnier”), Werktitel (“Séraphita”), Werkstoffe (“China-Papier”), sowie Ortsnamen (“Sevilla”) werden mit verschiedenen Tags (I-PERS, I-MISC, I-LOC) gekennzeichnet. Wie sich zeigt, werden jedoch nicht alle Eigennamen aufgefunden (etwa “Balzac, Honoré de”), was v.a. darauf zurückzuführen ist, dass hier zunächst nur das für das Deutsche verfügbar NER-Modell zum Einsatz kam. In der weiteren Projektlaufzeit muss die Erkennungsrate durch eine Erweiterung und Modifikation der vorhandenen Modelle verbessert werden, damit die Daten möglichst präzise und vollständig auswerten werden können.

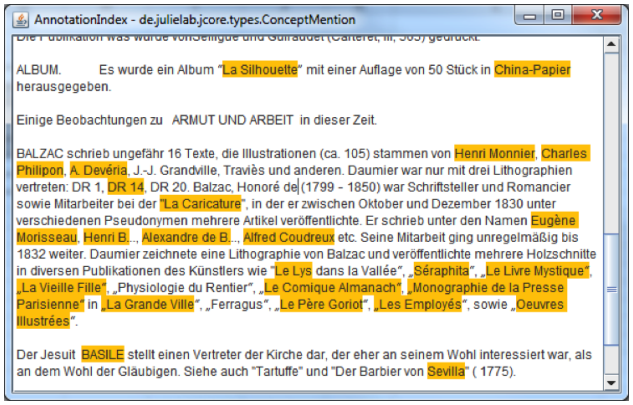


Abb. 4: Exemplarisches Ergebnis der Anwendung eines verfügbaren Standard-Modells zur Named Entity Recognition.

Zusammenfassung und Ausblick

Das beschriebene Vorgehen wird derzeit exemplarisch an ausgewählten Datensätzen entwickelt, um anschließend auf den gesamten Bildpool des prometheus-Bildarchivs angewendet zu werden. Ein wesentliches Ziel des Projekts ist es, eine größtmögliche Automatisierung in der Thesauruserstellung zu erreichen. Das hierfür vorgesehene kombinierte Ähnlichkeitsmaß ist flexibel erweiterbar. So können zum einen zusätzliche Informationen aus externen Quellen herangezogen werden, etwa indem weitere, digital vorliegende kunsthistorische Texte (z.B. das Reallexikon zur Deutschen Kunstgeschichte, RDK, siehe http://www.rdklabor.de), Ausstellungs- und Auktionskataloge (z.B. Getty Art and Architecture, siehe http://www.getty.edu/research/tools/vocabularies/aat/ , UB Heidelberg http://artsales.uni-hd.de) oder auch Wikipedia analysiert und klassifiziert werden und mittels des erstellten Thesaurus mit den Datensätzen in prometheus verknüpft werden. Zum anderen soll das Ähnlichkeitsmaß auch durch komplementäre (z.B. optische) Verfahren des Bildvergleichs erweitert werden. So wurde bspw. zusammen mit der Computer Vision Group Heidelberg bereits ein Projekt zur automatischen Bilderkennung angestoßen (siehe Bell/Dieckmann 2015). Durch die Kombination verschiedener Methoden der Ähnlichkeitsberechnung zu einem gemeinsamen, multidimensionalen Ähnlichkeitsmaß ist der hier vorgeschlagene Ansatz in hohem Maße adaptierbar für vergleichbare Anwendungen. Die im Projekt

erarbeitete Vorgehensweise ist somit auf weitere Metadatenpools kulturhistorischer Inhalte übertragbar und dank der Automatisierung beliebig skalierbar.

Bibliographie

Bell, Peter / Dieckmann, Lisa (2015): „Die Kunst als Ganzes. Heterogene Bilddatensätze als Herausforderung für die Kunstgeschichte und die Computer Vision“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* http://dhd2016.de/boa.pdf#118 [letzter Zugriff 23.11.2016].

Bell, Peter / Dieckmann, Lisa / Ommer, Björn / Takami, Masato (2015): *Passion Search. Prototype of an unrestricted image search of the crucifixion*. http://hci.iwr.uni-heidelberg.de/COMPVIS/projects/suchpassion/ [letzter Zugriff 23.11.2016]

Dieckmann, Lisa (2015): „prometheus – das verteilte digitale Bildarchiv für Forschung & Lehre e. V.“, in: Euler, Ellen / Hagedorn-Saupe, Monika/ Maier, Gerald/ Schweibenz, Werner/ Sieglerschmidt, Jörn (eds.): *Handbuch Kulturportale. Online-Angebote aus Kultur und Wissenschaft*. Berlin / Boston: DeGruyter 223–229.

Hermes, Jürgen (2012): *Textprozessierung: Design und Applikation*. Dissertation, Universität zu Köln. http://kups.ub.uni-koeln.de/id/eprint/4561 [letzter Zugriff 23. November 2016].

Neuefeind, Claes (2013): „The Digital Romansh Chrestomathy. Towards an Annotated Corpus of Romansh“, in: Zampieri, Marcos / Diwersy, Sascha (eds.), *Special Volume on Non-Standard Data Sources in Corpus-Based Research* (ZSM Studien 5). Aachen: Shaker 41–58.

Schwiebert, Stephan (2012): *Tesla - ein virtuelles Labor für experimentelle Computer- und Korpuslinguistik*. Dissertation, Universität zu Köln. http://kups.ub.uni-koeln.de/id/eprint/4571 [letzter Zugriff 23. November 2016].

Das „Was-bisher-geschah“ von KOLIMO. Ein Update zum Korpus der literarischen Moderne

Herrmann, J. Berenike
bherrma1@gwdg.de
Universität Göttingen, Deutschland

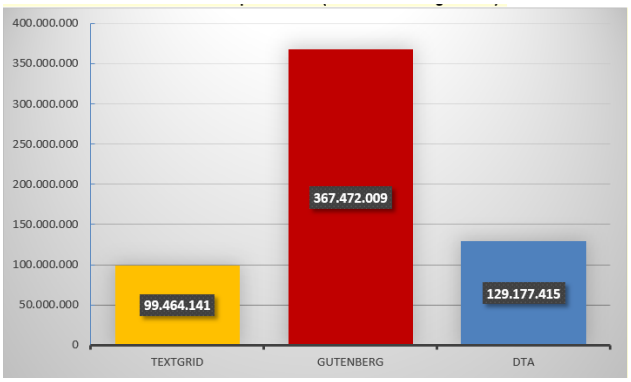
Lauer, Gerhard
Gerhard.Lauer@phil.uni-goettingen.de
Universität Göttingen, Deutschland

Der vorgeschlagene Beitrag dokumentiert den Fortschritt beim Aufbau unseres digitalen Korpus der literarischen Moderne (KOLIMO), das im Herbst 2016 in der Beta-Version veröffentlicht werden soll (abrufbar unter https://kolimo.uni-goettingen.de/). Im Fokus des Beitrags stehen das Verfahren zur Aufbereitung der Texte (insb. Format und Metadaten in TEI) und das linguistische Tagging (POS).

Als Teil des laufenden Projektes Q-LIMO (Quantitative Analyse der literarischen Moderne) ist KOLIMO ein repräsentatives und computerlinguistisch solide aufbereitetes Korpus von narrativen fiktionalen Erzähltexten der literarischen Epoche der Moderne. Um durch stratifiziertes Sampling Repräsentativität (verstanden als „extent to which a sample includes the full range of variability in a population“; vgl. Biber 1994) zu ermöglichen, umfasst das Korpus ein möglichst breites Spektrum der literarischen Moderne, verteilt über kanonische und nichtkanonische Texte. So wurden in das Korpus bislang ca. 596.000.000 Wörter aus frei zugänglichen Repositorien importiert (s. Abbildung 1).

Abbildung 1

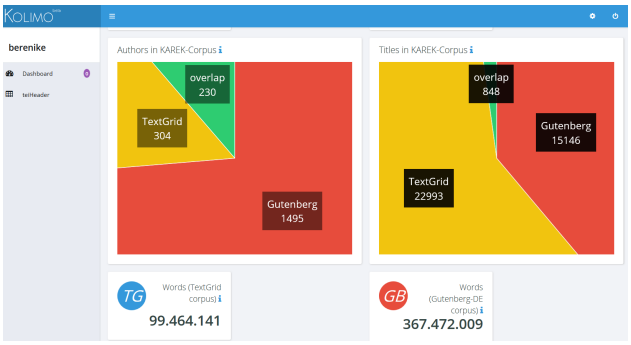
Gesamtanzahl Wörter aus den drei Hauptressourcen (Zwischenstand August 2016)



Die Datenbank umfasst so neben Texten aus TextGrid und Gutenberg-DE (s. Abbildung 2) und dem DTA auch eine wachsende Zahl von Retrodigitalisaten. Das Sampling ist nicht zuletzt dadurch beeinflusst, dass KOLIMO auch das Kafka/Referenzkorpus (KAREK) beinhaltet, welches zum Ziel hat, Kafkas Texte und Texte, die Kafkas Schreibprozess beeinflusst haben könnten, möglichst umfangreich abzubilden (vgl. Herrmann / Lauer 2016a,b).

Abbildung 2

Screenshot KOLIMO-WebApp: Anzahl Wörter, Autoren und Einträge aus TextGrid & Gutenberg-DE (ohne DTA und andere Quellen, Stand August 2016)



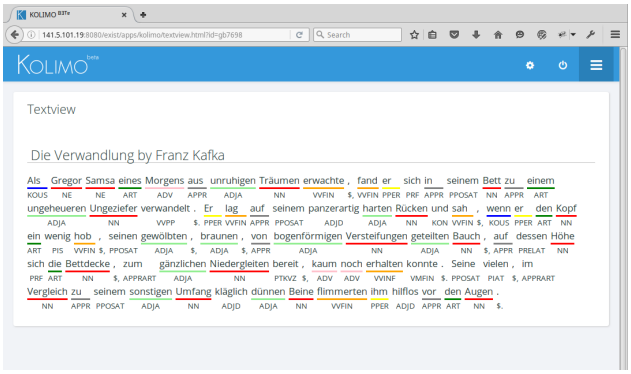
Um philologischen Ansprüchen an den editorischen Status literarischer Texte und die Abbildung von Epochen sowie Gattungskonzepten zu genügen, war eine hohe Genauigkeit und Konsistenz bei der informatischen Vorverarbeitung Textmarkup (XML-TEI) inklusive der Metadaten (Autor, Entstehungszeitpunkt und Gattung) besonders wichtig. Gerade die Auszeichnung der genannten Metadaten stellt eine Schnittstelle zwischen den informatischen und philologischen Dimensionen unseres Projektes dar: so sind Metadaten (a) die unabhängigen Variablen unserer stilistischen Analyse und (b) variieren in den von uns importierten Korpus-Ressourcen stark in qualitativer und quantitativer Hinsicht (Fehler, missing entries, unterschiedliche Ontologien). Der vorgeschlagene Beitrag wird so erstens einen kurzen Einblick in unsere Vorgehensweise geben, wobei Kriterien der Nachhaltigkeit berücksichtigt werden:

- Strategien der Textextraktion nach Genre-Kriterien unter Nutzung bestehender Metadatenschemata (ausgeschlossen wurden z.B. alle Texte, deren Metadaten sie als dramatisch und lyrisch ausflaggten, sowie Texte, die keine Absätze [without (tei:p)] enthielten);
- ein transparenter Workflow zur Korpusauszeichnung (internes eXist Webinterface);
- Anwendung eines standardisierten Text-Markups (u.a. Transformation der TextGrid und Gutenberg Header in das DTA-Basisformat TEI);
- Strategien der konsistenten Implementierung und Verbesserung von Metadatenschemata (Ineinandergreifen von händischen und skriptgestützen Workflows, wie Recherche zu [Erst-]Erscheinungsdaten bei missing entries, Zusammenführung der unterschiedlichen Gattungschemata, Überprüfung und ggf. Zuweisung von GNDs für Autoren);
- die nachhaltige Veröffentlichung des Korpus auf einem eigenen Server mit standardisierten Datenschnittstellen;
- Datenbankabbild (nonpublic) zur Langzeitarchivierung.

Zweitens wird der Beitrag unser Vorgehen bezüglich der linguistischen Anreicherung zusammenfassen: Unter der Annahme, dass Stil quantitativ beschreibbar ist (vgl. Herrmann / van Dalen-Oskam / Schöch 2015), und dass Wortarten verlässliche Indikatoren für Register und Genrevariation sind (vgl. z.B. Biber / Conrad 2009), haben wir uns für die linguistische Annotation auf POS (STTS Tagset; vgl. Schiller / Teufel / Thielen 1995) entschieden. POS sind im Vergleich mit anderen Variationsmarkern durch eine relativ akkurate automatische Annotation besonders praktikabel. Das Webinterface liefert variablen Zugriff auf die annotierten Daten, u.a. eine Volltextansicht (siehe Abbildung 3); geplant sind zur Veröffentlichung die Exportierbarkeit in .csv-Files und TCF-Format.

Abbildung 3

Screenshot KOLIMO WebApp Textview POS-Tagging



Zwar liefern bereits trainierte Modelle von einigen Taggern (z.B. TreeTagger) eine gute Genauigkeit für das gegenwärtige Standarddeutsch, angewendet auf ältere Sprachstufen oder vom Standarddeutschen abweichende Register wie „Literatur“ sinkt die Genauigkeit jedoch. Ein bereits auf POS annotiertes Korpus ist das Deutsche Textarchiv (DTA, Berlin-Brandenburgische Akademie der Wissenschaften 2016), ein Referenzkorpus für das Deutsche, das sowohl historische Sprachstufen als auch das Register „Literatur“ enthält. Die POS-Annotation baut hier auf fehlertoleranten linguistischen Analyse historischer Texte auf und verwendet ein Tool zur Morphologisierung (Jurish 2012), ist allerdings hinsichtlich ihrer Qualität noch nicht umfassend evaluiert worden. Ausgehend von diesem Datensatz haben wir zwei Strategien verfolgt: (1) Ein epochensensitives POS-Tagging, das verschiedene Tagger auf dem Datensatz des DTA, aber auf unterschiedlichen literarischen Epochen trainiert (vgl. Paluch et al. in Vorbereitung); (2) eine Überprüfung der Qualität der DTA-POS-Tags durch quantitative und qualitative Verfahren.

In Strategie (1) machen wir uns zunutze, dass Annotationsgenauigkeit erhöht werden kann, wenn Tagger auf verschiedene Register/ Sprachstände trainiert und diese trainierten Modelle dann auf noch nicht trainierte Texte des gleichen Registers angewendet werden (vgl. Giesbrecht / Evert). Für KOLIMO haben wir u.a. den TreeTagger (vgl. Schmid 1994), Perceptron (vgl. Rosenblatt 1958) und MarMoT (vgl. Müller / Schmid / Schütze 2013) verwendet. Durch die Wahl unterschiedlicher Tagger soll gewährleistet werden, dass die Genauigkeit der POS-Annotation maximiert werden kann, indem nur derjenige Tagger mit den besten Ergebnissen pro Register verwendet wird. Die Auswahl der Tagger basierte einerseits darauf, dass sie unterschiedliche Prinzipien benutzen: So funktioniert der TreeTagger nach dem Hidden Markov Model (HMM, vgl. Baum / Petrie 1966), MarMot nach dem Prinzip der Conditional Random Fields (DRF, vgl. Hammersly / Clifford 1971) und Perceptron nach dem neuronaler Netzwerke. Der Grund für die Wahl des TreeTaggers war zudem seine Prävalenz in der Forschungsliteratur, die nicht zuletzt durch gute Ergebnisse begründet scheint (vgl. Dipper 2012; Giesbrecht / Evert 2009). In einem ersten Schritt (vgl. Paluch et al. in Vorbereitung) wurden hier bereits getaggte Texte aus dem DTA in fünf Epochen geordnet. Neben der Moderne umfassten diese zu Vergleichszwecken auch Barock, Aufklärung, Romantik, und Realismus. Für die Einteilung der Epochen in Zeitperioden sowie der Einteilung von Autoren zu bestimmten Epochen wurden einschlägige Literaturgeschichten zu Rate gezogen (u.a. Beutin 2001; Jørgensen / Bohnen / Øhrgaard 1990; Meid 2009; Schulz 2000; Sprengel 1998, 2004). Anschließend wurden die Tagger auf jeweils eine Epoche trainiert, indem die Texte randomisiert in Trainings- und Evaluationstexte getrennt wurden und eine k-fold cross validation (vgl. Witten / Elbe 2005) für jeden Tagger durchgeführt wurde. Die Ergebnisse (vgl. auch Paluch et al. in Vorbereitung) weisen auf eine gute Genauigkeit insbesondere von Perceptron hin, müssen aber unter dem Vorbehalt betrachtet werden, dass der Status des DTA als Goldstandard für POS-Tagging noch fraglich ist.

Hier setzen wir mit Strategie (2) an, mit der wir zunächst für alle POS-Tags Übereinstimmung und Abweichung (Matches und Missmatches) des Outputs des Tree-Taggers und MarMots mit dem DTA-Datensatz vergleichen. Aufbauend auf diese quantitative Überprüfung der einzelnen Tag-Zuweisung evaluieren wir zudem händisch Stichproben der Nichtübereinstimmungen in der Annotation der einzelnen Tags.

Unsere quantitative Überprüfung ergibt eine generelle Übereinstimmung mit dem DTA-Datensatz in POS-Tags für den TreeTagger und den Marmot Tagger von jeweils 80%. Die generelle Übereinstimmung zwischen den Tags des TreeTaggers und denen des MarMot Taggers hingegen liegt bei 0.78%.

Tabelle 1 zeigt Ergebnisse aus der Analyse der Übereinstimmungen (Matches) und Abweichungen (Missmatches) bei der POS-Tagzuweisung von TreeTagger (TT) und MarMot (MM) im Vergleich mit den Tags des DTA. Abgebildet sind hier solche Fälle pro POS-Tag, in denen TT und MM übereinstimmen, aber vom DTA abweichen. Die Tabelle listet die elf POS-Tags, die (von TT und MM gemeinsam) die proportional den höchsten Anteil der Abweichung vom DTA ausmachen.

Tabelle 1 Abweichung zu POS-Tags des DTA (Übereinstimmung MM und TT)

POS-Tag*	Häufigkeit	Rel. Häufigkeit
NE	1444048	0.12
NN	1443795	0.12
VVFIN	1326081	0.11
ADJA	1309006	0.11
ADJD	741903	0.06
ADV	618465	0.05
VAFIN	582791	0.05
FM.la	404341	0.03
PPOSAT	397465	0.03
APPR	362774	0.03
PDAT	255896	0.02

*STTS Tagset

Aufbauend auf diesen Daten wird im nächsten Schritt die tatsächliche Qualität der bereits vorhandenen DTA-Tags für den Datensatz der literarischen Texte evaluiert. Auf der Grundlage von randomisiertem Sampling verbessern wir die POS-Annotationen bei tatsächlichen Fehlern händisch, um in der Folge u.a. eigene Sprachmodelle für unser spezifisches Korpus narrativer Texte zu trainieren. So soll schließlich unter Nutzung vorhandener Ressourcen ein Silber- oder sogar Goldstandard für das POS-Tagging historischer literarischer Texte des Deutschen erreicht werden.

KOLIMO wird in der Beta-Version zur Tagung veröffentlicht (s. https://kolimo.uni-goettingen.de) und so der Forschungsgemeinschaft zur Verfügung gestellt. Es soll eine hypothesengetriebene, aber auch explorative, quantitative Stilistik ermöglichen (vgl. Herrmann eingereicht); zum Zeitpunkt der Tagung sind erste Ergebnisse zur stilistischen Variation der literarischen Moderne zu erwarten (vgl. schon Herrmann / Lauer / Mattner 2016).

Gleichzeitig planen wir eine detaillierte Dokumentation der Arbeitsschritte zu veröffentlichen, die ähnlichen Projekten als Leitfaden zur Verfügung zu stehen soll. Unser Projekt dokumentiert in seinem gegenwärtigen Status Entscheidungen auf verschiedenen konzeptionellen, analytischen und prozeduralen Ebenen. Es zeigt, dass der Aufbau eines digitalen literarischen Korpus, das den synchronen und diachronen quantitativen Vergleich einer Schwerpunktepoche erlauben soll, bei Weitem keine triviale Aufgabe darstellt. So wurde zum Beispiel deutlich, wie Hypothesen zur Konstitution von Epochen, Autorschaft und Gattungen die Korpuskompilation steuern – und deshalb auf einer möglichst präzisen Modellierung der zugrundeliegenden textwissenschaftlichen Theorien fußen sollten. Gleichzeitig sind Metadaten (u. a. Autor, Titel, Publikationsdatum, Publikationsort, Gattung) und linguistische Parameter (wie POS) gerade die Ansatzpunkte, an denen philologische Fragestellungen in präzise und praktikable Kategorien umgewandelt werden können. Nicht zuletzt deshalb sollten literarische Daten in flexiblen Architekturen gespeichert werden, die zusätzliche Annotationsebenen zulassen – denn hermeneutische Erkenntnisprozesse stellen eine erwachsene Stärke der Geisteswissenschaften dar, die auch im digitalen Zeitalter einen explizit modellierten Platz einnehmen muss.

Bibliographie

Baum, Leonard E. / Petrie, Ted (1966): „Statistical inference for probabilistic functions of finite state markov chains“, in: *The annals of mathematical statistics 37* (6) :1554–1563.

Berlin-Brandenburgische Akademie der Wissenschaften (2016): *Deutsches Textarchiv*. http://www.deutschestextarchiv.de/ [letzer Zugriff 24. Mai 2016].

Beutin, Wolfgang (2001): *Deutsche Literaturgeschichte: von den Anfängen bis zur Gegenwart*. Stuttgart: Metzler.

Biber, Douglas / Conrad, Susan (2009): *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Dipper, Stefanie (2012): „Morphological and part-of-speech tagging of historical language data: A comparison“, in: *Workshop on Annotation of Corpora.* http://www.coli.uni-saarland.de/conf/ACRH10/slides/dipper.pdf .

Gaede, Friedrich (1971): *Humanismus, Barock, Aufklärung: Geschichte der deutschen Literatur vom 16. bis zum 18. Jahrhundert*. Bern: Francke Verlag.

Giesbrecht, Eugenie / Evert, Stefan (2009): „Is part-of-speech tagging a solved task? An evaluation of pos taggers for the German web as corpus“, in: *Proceedings of the fifth Web as Corpus Workshop* 27–35.

Hammersley, John M. / Clifford, Peter (1971): *Markov fields on finite graphs and lattices*. http://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf .

Herrmann, J. Berenike (eingereicht): „In test bed with Kafka. Introducing a mixed-method approach to digital stylistics“, in: Chambers, Sally / Jones, Catherine / Kestemont, Mike / Koolen, Marijn / Zundert, Joris van (Eds.). *Special*

Issue DHBenelux 2015, Digital Humanities Quarterly.

Herrmann, J. Berenike / Lauer, Gerhard (2016a): „KAREK: Building and Annotating a Kafka/Reference Corpus“, in: *DH2016: Conference Abstracts*.

Herrmann, J. Berenike / Lauer, Gerhard (2016b): „Aufbau und Annotation des Kafka/Referenzkorpus“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung*.

Herrmann, J. Berenike / Lauer, Gerhard / Mattner, Cosima (2016): *Measuring Kafka's Diaries*. A Psychostylistic Approach International Society for the Empirical Study of Literature and Media (IGEL), Chicago, USA.

Herrmann, J. Berenike / van Dalen-Oskam, Karina / Schöch, Christof (2015): „Revisiting Style, a Key Concept in Literary Studies“, in: *Journal of Literary Theory* 9 (1): 25–52.

Jørgensen, Sven Aaage / Bohnen, Klaus / Øhrgaard, Per (1990): *Aufklärung, Sturm und Drang, frühe Klassik: 1740 - 1789*. (Boor, Helmut de / Newald, Richard, eds.). München: Beck.

Jurish, Bryan (2012): *Finite-state Canonicalization Techniques for Historical German*. PhD, Universität Potsdam.

Manning, Christopher D. / Raghavan, Prabhakar / Schütze, Heinrich (2008): *Introduction to information retrieval* 1. Cambridge: Cambridge University Press.

Meid, Volker (2009): *Die deutsche Literatur im Zeitalter des Barock: vom Späthumanismus zur Frühaufklärung: 1570 - 1740*. (Boor, Helmut de / R. Newald, Richard, eds.) ([Neuausg.].). München: Beck.

Müller, Thomas / Schmid, Helmut / Schütze, Hinrich (2013): „Efficient higher-order CRFs for morphological tagging“, in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Nekula, Marek (2003): „Franz Kafkas Deutsch“, in: *Linguistik online* 13 (1) https://bop.unibe.ch/linguistik- online/article/view/879/1533 .

Paluch, Markus / Rotari, Gabriela / Steding, David / Weß, Maximilian / Moritz, Maria (in Vorbereitung): *Non-static analysis of part-of-speech tagging of historical German texts*.

Rosenblatt, Frank (1958): „The perceptron: a probabilistic model for information storage and organization in the brain“, in: *Psychological Review* 65 (6): 386.

Schiller, Anne / Teufel, Simone / Thielen, Christine (1995): „Guidelines für das Tagging deutscher Textcorpora mit STTS“, in: *Manuscript, Universities of Stuttgart and Tübingen*. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf .

Schmid, Helmut (1994): „Probabilistic part-of-speech tagging using decision trees“, in: *Proceedings of the international conference on new methods in language processing* 12: 44–49.

Schulz, Gerhard (2000): *Das Zeitalter der Französischen Revolution: 1789 - 1806*. (Boor, Helmut de / Newald, Richard, eds.) (2., neubearb. Aufl.). München: Beck.

Sprengel, Peter (1998): *Geschichte der deutschsprachigen Literatur 1870 - 1900: von der Reichsgründung bis zur Jahrhundertwende*. (Boor, Helmut de / Newald, Richard, eds.). München: Beck.

Sprengel, Peter (2004): *Geschichte der deutschsprachigen Literatur 1900 - 1918: von der Jahrhundertwende bis zum Ende des Ersten Weltkriegs*. (Boor, Helmut de / Newald, Richard, eds.). München: Beck.

Witten, Ian H. / Elbe, Frank (2005): *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann Publishers.

Datenmodellierung und -visualisierung mit Graphdatenbanken. Konzepte und Erfahrungen anlässlich des Relaunches der Bilddatenbank REALonline

Matschinegg, Ingrid
ingrid.matschinegg@sbg.ac.at
Universität Salzburg, Österreich

Nicka, Isabella
isabella.nicka@sbg.ac.at
Universität Salzburg, Österreich

Mit der Thematik der digitalen Nachhaltigkeit sind Gedächtnisinstitutionen, die sich die umfassende digitale Sicherung und Erschließung des Kulturerbes zum Ziel setzen, gleichermaßen konfrontiert wie Forschungsprojekte, bei denen digitale Daten generiert, computergestützt ausgewertet und in Forschungsdatenbanken zugänglich gemacht werden. Dabei liegt

das Hauptaugenmerk nicht nur auf der Langzeitsicherung und Archivierung; die größeren Herausforderungen stellen sich vor allem bei den Aktualisierungen der Datenmodelle, Analysemethoden und Präsentationsformen, um die neuen Werkzeuge der Digital Humanities bestmöglich einsetzen zu können. Die Überlegungen im Vorfeld derartiger Relauncharbeiten bewegen sich erfahrungsgemäß zwischen vorsichtiger Adaptierung und radikalem Umbau der vorliegenden Datenarchitektur. Dass dabei alle vorhandenen Informationen verlustfrei übertragen werden sollen, versteht sich von selbst. Der folgende Beitrag möchte die Transformation von Daten eines Langzeitprojekts in eine neue Datenarchitektur für eine Bilddatenbank vorstellen, bei der eine Graphdatenbank zum Einsatz kommt:

Am Institut für Realienkunde des Mittelalters und der frühen Neuzeit (IMAREAL), einem interdisziplinär ausgerichteten Forschungsinstitut, das Teil der Universität Salzburg ist, wird die materielle Kultur des Mittelalters und der frühen Neuzeit untersucht. Bildquellen bilden dabei neben Schriftquellen und überlieferten Objekten die Grundlagen der Analysen. Mit dem Aufbau der Bilddatenbank REALonline wurde am IMAREAL in den 1970ern auf der Grundlage der von Manfred Thaller speziell für die Anforderungen der historischen Grundwissenschaften entwickelten Datenbanksysteme begonnen – zunächst *Descriptor* und in weiterer Folge *Κλειώ* (Thaller 1980 u. 1989). Der Datenbestand von REALonline wurde seither und wird weiterhin kontinuierlich erweitert, damit dargestellte Dinge und ihre Kontexte erforscht werden können. Die Datenbank ist seit 2002 unter http://tethys.imareal.sbg.ac.at/realonline online verfügbar (Matschinegg 2004). Anhand der Datenbank ist es möglich, die Bedeutung und Funktion der materiellen Kultur im Bilddiskurs zu untersuchen: Welche Objekte waren zu welchen Zeiten in welchen Gesellschaften und Kontexten als visuelle „Requisiten“ gegenwärtig oder vor- und damit auch darstellbar? Wie wurden Dinge im Bild verhandelt und welche Rolle nehmen sie innerhalb von ins Bild überführten Narrativen ein?

Um diese Fragen beantworten zu können, wurde am IMAREAL entschieden, neben den Metadaten zum Werk bzw. Bildträger systematisch *alle* im Bild dargestellten Elemente auszuzeichnen (Abb. 1). Im Gegensatz zu anderen Bilddatenbanken wird das Dargestellte nicht nur mit einigen wenigen Schlagwörtern erfasst. Diesem Umstand ist es zu verdanken, dass die in REALonline erhobenen Daten sowohl im Rahmen von interdisziplinären Forschungen zur materiellen Kultur ausgewertet werden können, als auch in unterschiedlichen geisteswissenschaftlichen Untersuchungskontexten und für Kulturerbedokumentationen eine wertvolle Ressource darstellen.

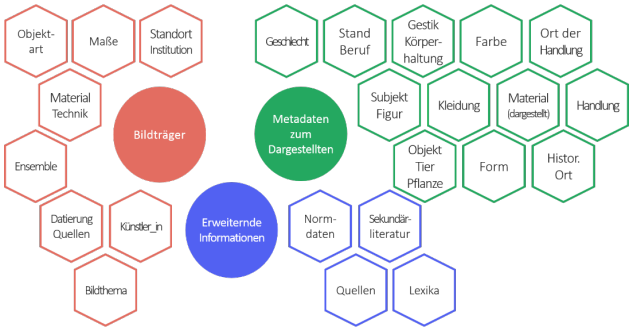


Abb. 1: Erfassungsschema der Metadaten in REALonline

Im Modell für die Erfassung der dargestellten Entitäten im Bild werden folgende Informationen erhoben: Für Subjekte werden neben dem Subjektnamen die Kategorien Geschlecht, Beruf bzw. Stand und Gestik erfasst. Bei Objekten werden der Objektname und die Informationen zu Farbe, Material und Form erhoben. Weiters wird die Struktur dieser Metadaten zu den Bildinhalten festgehalten und kann damit im Rahmen von Analysen abfragbar gemacht werden: Direkte Subjekt-Objekt- bzw. Objekt-Objekt-Relationen werden erfasst, um am Körper getragene bzw. von Figuren gehaltene Objekte zu dokumentieren oder einen Bezug zwischen einzelnen dargestellten Dingen (etwa ein auf einem Tisch stehender Krug) in den Daten abbilden zu können. Darüber hinaus können sowohl Körperteile als auch Teile von Objekten als Metadaten zum Dargestellten gespeichert werden (Jaritz 1993: 23-43).

Graphdatenbanken eignen sich u.a. besonders dafür, die vielfältigen Beziehungen zwischen Personen oder Personen und Gegenständen bzw. Ereignissen sowie auch zwischen den Dingen untereinander möglichst flexibel abzubilden und sind nun auch in der historischen Forschung im Kommen; als Software wird oft Neo4j eingesetzt (Raspe 2014, Kaufmann & Andrews 2015, Kuczera 2015). Die verzweigte Struktur der erfassten Metadaten in REALonline ist einer der Hauptgründe, warum für die neue Datenarchitektur ein property-graph-Modell gewählt wurde. Ein weiterer Leitgedanke war, dass das Beziehungsnetz von Subjekten, Objekten und Handlungen in mittelalterlichen

und frühneuzeitlichen Bildern anhand des Modells eines verzweigten Graphen besser veranschaulicht werden kann als in einer langen Liste mit Metadaten. Die Graphdatenbank bietet im Fall von REALonline sowohl bei der Präsentation für die Nutzer_innen im Frontend, für die Abfrage und Darstellung der Abfrageergebnisse als auch für die Eingabe der Daten im Backend (siehe Abb. 2) eine Verbesserung der Usability gegenüber dem zuvor verwendeten hierarchischen Datenbankmodell.

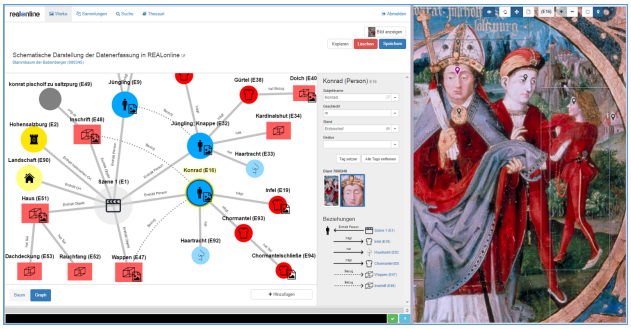


Abb. 2: Screenshot des Graphen zum Dargestellten im Bild (Backend)

In unserem Projekt hat sich die Kombination von Neo4j für die Modellierung und Abfrage der „beziehungsrelevanten“ Daten mit einer NoSQL-Mongo-Dokumentendatenbank angeboten (Abb. 3). Diese Lösung baut einerseits auf dem in der Praxis bereits bewährten softwareseitigen Ineinandergreifen bei der semantischen Transformation der Informationen auf und bietet gleichzeitig die Möglichkeit zur Speicherung und Abfrage von werkgeschichtlich wie auch projektgeschichtlich relevanten Informationen zu den einzelnen Bilddokumenten, die im Verlauf dieses Langzeitprojektes erhoben wurden und laufend erweiterbar bleiben sollen.

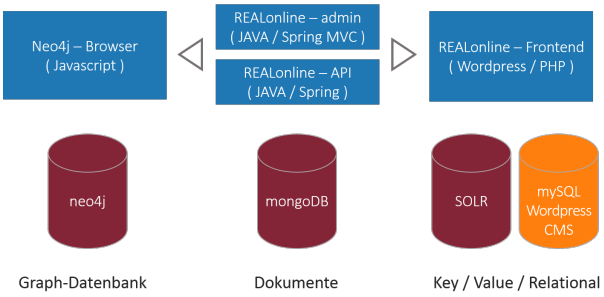


Abb. 3: Datenarchitektur von REALonline

Im Vortrag möchten wir aber auch auf die Herausforderungen hinweisen, denen wir uns im Zuge des Entwurfs der Datenarchitektur von REALonline stellen mussten: So war etwa in der Struktur des hierarchischen Datenmodells die Information zur dargestellten Handlung auf derselben Ebene angesiedelt wie die Entitäten Subjekt und Objekt. Beim Datenexport aus der bis dato verwendeten *Κλειώ* -Datenbank und dem Import in Neo4j konnte – nachdem in diesem Fall keine automatisierten Zuweisungen der Handlung zu Personen bzw. Objekten möglich waren – dieser Umstand nur in das neue Datenmodell mitübernommen werden. Mit der Entscheidung für eine Graphdatenbank ist dennoch gewährleistet, dass in einem weiteren Schritt Informationen, wie jene zur dargestellten Handlung, statt in Knoten in die Kanten des Graphen gelegt werden können und damit die Struktur von RDF-Triples (Subjekt-Prädikat-Objekt) bekommen.

Aufgrund der zeitintensiven Datenerhebung war ein wichtiger Aspekt des Relaunchs, die Dateneingabe so effizient wie möglich zu gestalten. Der Beitrag wird die gefundene Lösung präsentieren. Langfristig gesehen sollte versucht werden, den Zeitaufwand für die Erhebung von Metadaten zu den auf historischen Bildern dargestellten Elementen zu minimieren. Daher möchten wir die in REALonline während mehr als 40 Jahren erhobenen Informationen als Trainingsdaten in transdisziplinäre Projekte zwischen den Geisteswissenschaften und der Computer Vision – insbesondere zur (semi-)automatisierten Bilderkennung – einbringen, so dafür Fördermittel eingeworben werden können.

Mit dem Relaunch von REALonline kann die Menge der erhobenen Metadaten zum im Bild Dargestellten (aktuell sind innerhalb von 23316 Datensätzen 1.165562 Begriffe dazu erfasst) besser zugänglich gemacht werden: Abfragen der Graphdatenbank und Visualisierungen (z.B. mit Software wie *gephi– The Open Graph Viz Platform* oder *yEd graph editor*) dieser Ergebnisse können komplexe Zusammenhänge innerhalb der Bilddetails aufdecken oder Aufschlüsse zu Mustern sowie „Ausreißern“ in unterschiedlichen Samples liefern, die nicht nur als Resultate statistischer Auswertungen verstanden werden sollen, sondern vor allem dazu dienen können, neue Fragen in der (interdisziplinären) Forschung anzustoßen. Beispielsweise wurde in 154 Datensätzen das Bildthema „Geißelung Christi“ erfasst. Bei der Erschließung dieser Datensätze wurden wiederum 516 Objekte verzeichnet, die von Figuren im Bild in der Hand gehalten werden. Die Visualisierung (Abb. 4) beschränkt sich auf jene Objekte, die nur einmal vorkommen (gelb) und die ihnen

übergeordneten Thesauruskategorien (grün, dunkelrot). Während die meisten Objekte dem gängigen Narrativ „Geißelung“ zugeordnet werden können, sind die Objekte *Münze* und *Geldbeutel* nur über einen Konnex zum mittelalterlichen Drama erklärbar (Nicka 2014, 280–282): Die Darstellung einer Bezahlung der Geißler Christi, die nur auf einem Flügelaltar im niederösterreichischen Pöggstall im Bild festgehalten wurde, kennen wir ansonsten nur aus Passionsspielen, wo jüdische Protagonisten negativ gekennzeichnet werden, indem sie den Gerichtsknechten Geld geben, um besonders fest mit den Ruten zuzuschlagen (siehe auch Abb. 2).

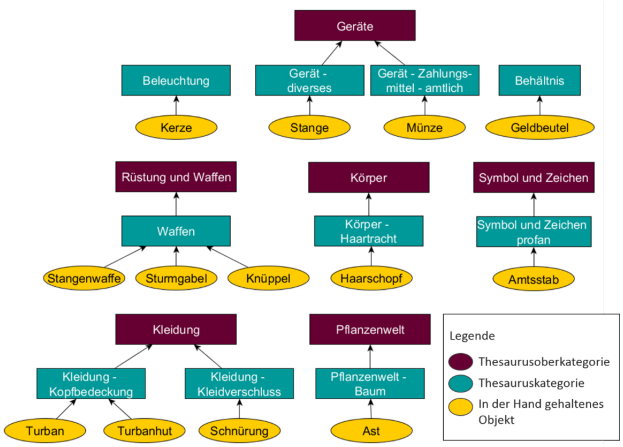


Abb. 4: Visualisierung der dargestellten Objektbegriffe und Körperbezeichnungen aus den Bildbeschreibungen in ihrer Zuordnung zur jeweiligen Thesauruskategorie

Abschließend bleibt zu erwähnen, dass sich mit der Notwendigkeit, eine gut eingeführte, aber in ihren technischen Funktionalitäten nicht mehr zeitgemäße Bilddatenbank zu modernisieren, auch die Chance zur besseren Nutzung der umfangreichen Datenbestände verbinden lässt. Im Zuge der Relaunchvorarbeiten haben wir die Konzepte und Lösungsansätze aufgegriffen, die gegenwärtig in den Digital Humanities diskutiert und getestet werden. Wir haben die Umsetzung in enger Zusammenarbeit mit den Grazer Entwicklerfirmen complement.at und zedlacher.net realisiert. Der Beitrag gibt einen knappen Überblick über die wichtigsten Entscheidungsfindungsprozesse sowie die Schwierigkeiten und Potentiale, die bei der Überführung in die neue Datenarchitektur und die gewählte Frontend-Lösung entstanden sind. Der Aspekt der Nachhaltigkeit hat dabei von Anfang an eine große Rolle gespielt; sowohl bei der Erhaltung aller vorhandenen Informationen als auch bei der nachhaltigen Nutzbarkeit der erhobenen Daten. So ist die Zitierbarkeit der Daten über einen PID (persistent identifier) mit einem handle gewährleistet und die Metadaten werden mit einer *Creative Commons by-nc-sa* 4.0-Lizenz zur Verfügung gestellt. Die neue Online-Version von REALonline wird gegenwärtig getestet und optimiert und 2017 freigeschaltet.

Bibliographie

Jaritz, Gerhard (1993): *Images: A Primer of Computer-Supported Research with Κλειώ IAS*. Halbgraue Reihe zur historischen Fachinformatik A 22. St. Katharinen: Scripta Mercaturae Verlag.

Kaufmann, Sascha / Andrews, Tara Lee (2016): „Bearbeitung und Annotation historischer Texte mittels Graph-Datenbanken am Beispiel der Chronik des Matthias von Edessa“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 176–178 http://dhd2016.de/boa.pdf [letzter Zugriff 20. August 2016].

Kuczera, Andreas (2015): „Graphdatenbanken für Historiker. Netzwerke in den Registern der Regesten Kaiser Friedrichs III. mit neo4j und Gephi“, in: *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte*, 5. Mai 2015, http://mittelalter.hypotheses.org/5995 (ISSN 2197-6120) [letzter Zugriff 20. August 2016].

Matschinegg, Ingrid (2004): „REALonline – IMAREAL's Digital Image-Server“, in: *[Enter the Past]. The E-way into the Four Dimensions of Cultural Heritage. CAA 2003 | Computer Applications and Quantitative Methods in Archaeology | Proceedings of the 31st Conference*, Vienna, Austria, April 2003 (BAR International Series 1227). Oxford: archaeopress, 214-216.

Nicka, Isabella (2014): „Interfaces. Berührungszonen von Transzendenz und Immanenz im spätmittelalterlichen Sakralraum“, in: Meyer, Marion / Klimburg-Salter, Deborah (eds.): *Visualisierungen von Kult*. Wien / Köln / Weimar: Böhlau, 260-293, Abb. auf 438-444.

Nicka, Isabella (im Erscheinen): „REALonline–Explore and Find Out. Wohin führt das Digitale die Kunstgeschichte?“, Beitrag zum Tagungsband der vom 6.-8. Nov. 2015 in Wien abgehaltenen Konferenz *„Newest Art History“. Wohin geht die jüngste Kunstgeschichte?*

Raspe, Martin (2014): *Zuccaro. Ein modernes, konfigurierbares Informationssystem für die Geisteswissenschaften*. http://zuccaro.biblhertz.it/dokumentation/zuccaro [letzter Zugriff 20. August 2016].

Thaller, Manfred (1980): „Descriptor: Probleme der Entwicklung eines

Programmsystems zur computerunterstützten Auswertung mittelalterlicher Bildquellen“, in: *Europäische Sachkultur des Mittelalters*: Gedenkschrift aus Anlaß des 10jährigen Bestehens des Instituts für mittelalterliche Realienkunde Österreichs (Veröffentlichungen des Instituts für mittelalterliche Realienkunde Österreichs 4 / Sitzungsberichte der Akademie der Wissenschaften, Phil.-Hist. Klasse 374). Wien: Verlag der Österreichischen Akademie der Wissenschaften, 167–194.

Thaller, Manfred (1989): *Κλειώ*. Ein Datenbanksysten. Halbgraue Reihe zur historischen Fachinformatik B1. St. Katharinen: Scripta Mercaturae Verlag.

Datenvisualisierung als Aisthesis

Gius, Evelyn
evelyn.gius@uni-hamburg.de
Universität Hamburg, Deutschland

Kleymann, Rabea
rabea.kleymann@uni-hamburg.de
Universität Hamburg, Deutschland

Meister, Jan Christoph
jan.c.meister@gmail.com
Universität Hamburg, Deutschland

Petris, Marco
marco.petris@uni-hamburg.de
Universität Hamburg, Deutschland

Visualisierung als *Expansion*

Jede/r DH-Praktiker/in weiß: Computergestützte Forschung in den Geisteswissenschaften beginnt mit der Übersetzung relevanter Phänomene in digitale Daten. Seltener thematisiert wird dagegen, dass die digitale *Operationalisierung* (Moretti 2013) ein wichtiges Gegenstück am Ende des Forschungsprozesses hat: Die digitale Repräsentation des Untersuchungsgegenstandes wie das generierte Daten-Output müssen in eine nicht-algorithmisierte Form gebracht werden, um überhaupt sinnvoll von Menschen verstanden und weiter bearbeitet werden zu können. Für diesen Prozess der Rückübersetzung hat Goodings (2003) das Konzept der *Expansion* eingeführt. [1]

Derartige Expansionsverfahren systematisch zu beschreiben, ist in den Geisteswissenschaften [2] schwierig, besitzen Primär- und Metadaten [3] hier doch meist komplexe und teils sogar exponentiell expandierende Datenstrukturen. Mehr noch: Forschungsfragen, die den geisteswissenschaftlichen Erkenntnisprozess motivieren, sind typischerweise multidimensional; sie reflektieren Zustände, aber zugleich auch historische Abläufe und Bezüge; sie interagieren i.d.R. dynamisch mit dem Erkenntnisinteresse. Das macht Modellierungen wie Datenanalysen um ein Vielfaches aufwendiger – und dies nicht nur für den Computer, der die primären Eingabedaten verarbeitet und neue sekundäre Daten generiert, sondern vor allem für die Forscher/innen, die beide Datentypen wieder miteinander abgleichen und für ihren Erkenntnisprozess fruchtbar machen wollen.

Visualisierungen gelten heute disziplinübergreifend als probates Mittel der “Expansion” schwer überschaubarer Primär- und Sekundärdaten in intuitiv erfassbarer Form (Goodings 2003:281). In der disziplinspezifischen Perspektive ist allerdings zugleich zu fordern, dass den methodologischen Besonderheiten der *Humanities* Rechnung getragen wird. [4] Drei spezifische Merkmale sind hier zu berücksichtigen:

Geisteswissenschaftliche Verstehensprozesse
: sind grundsätzlich organisiert als fortlaufende, dynamische Iteration von empirisch-analytischen und theoretisch-modellierenden Operationen.

Geisteswissenschaftliche
: Interpretationsverfahren sind in der Regel ebenfalls nicht als unilineare ‘Auslegungen’ konzipiert, sondern wirken auf die Ausgangsdaten zurück, indem sie diese anreichern, relativieren oder rekonfigurieren.

Geisteswissenschaftliche Verstehensprozesse
: sind nicht nur in hohem Maße kontextsensitiv, sondern zudem reflexiv: Verstanden werden will nicht nur das je gegebene Untersuchungsobjekt, sondern verstanden werden sollen auch (und in Disziplinen wie der Literaturwissenschaft oder der Philosophie mitunter sogar primär) die Bedingungen und Möglichkeiten des Verstehensprozesses selbst.

Datenvisualisierungen, die als visuelle Expansionsverfahren *geisteswissenschaftlich* funktional sein sollen, müssen diese drei Prozessmerkmale in Form von methodischen wie technischen Spezifikationen abbilden. Sie müssen vor allen Dingen aber auch von einem übergreifenden Visualisierungskonzept angeleitet werden, das epistemologisch ausgerichtet ist und danach fragt, auf welche Art von Erkenntnis die Geisteswissenschaften eigentlich zielen.

Trotz der großen Vielfalt an bestehenden Visualisierungstools und Visualisierungsmetaphern gibt es allerdings bislang kein derartiges, theoretisch reflektiertes *Visualisierungskonzept*, das von einer Typologie der Forschungsfragen wie der methodischen Logik geisteswissenschaftlicher Forschungsprozesse her entworfen wäre. Digitale Visualisierungslösungen werden vielmehr von den Geisteswissenschaften unhinterfragt aus anderen Verwendungskontexten importiert (z. B. Kreisdiagramme, Verlaufskurven, Scatter Plots etc. aus der Statistik) oder bestenfalls als ein ‚irgendwie' erstaunlich funktionales Tool angenommen (z. B. Word Clouds). Das aber hat zur Folge, dass die Verstehensmöglichkeiten der Geisteswissenschaften von den je gewählten, vielfach aus evidenzzentriert verfahrenden Disziplinen übernommenen visuellen Metaphern determiniert werden. [5]

Zur Entwicklung einer visuellen Grammatik für hermeneutische Verstehensprozesse

Grundlagen: Interaktivität und methodische Passung

Gegenstand des Projekts "3DH – dreidimensionale dynamische Datenvisualisierung und Exploration für Digital Humanities-Forschungen" ist die Entwicklung und prototypische Implementierung eines solchen Konzepts der geisteswissenschaftlichen Datenvisualisierung. [6] Mit der 'dritten Dimension' ist dabei nicht primär die *räumliche* z-Achse gemeint, sondern grundsätzlicher die einer *konzeptionellen* 'Achse', die den methodologischen Erfordernissen der Geisteswissenschaften Rechnung trägt.

Grundlegendstes dieser Erfordernisse ist, die bildhafte Veranschaulichung von Daten konsequent bi-direktional zu denken. Die *interaktive Exploration* [7] geisteswissenschaftlicher Datenkomplexe ist deshalb methodische Leitidee für das im Projekt entwickelte Visualisierungskonzept. Konkret heißt dies: Der Bildschirm muss vom bloßen *Renderer* zum *Two Way Screen* werden, der nicht nur Daten und Datenstrukturen als visuelles Output darstellt, sondern umgekehrt auch deren interaktive Manipulation und Analyse ermöglicht. Damit wird der hermeneutischen Analyse- und Interpretationspraxis Rechnung getragen, in der Verstehen ein "produktives Verhalten" ist (Gadamer 1972:280).

Zweites Erfordernis ist, dass hermeneutisch funktionale Visualisierungen neben generischen Anforderungen auch die Besonderheiten geisteswissenschaftlicher Praxis in den *Einzeldisziplinen* berücksichtigen. Für deren je spezifische Datentypen und Modi der Datenaggregation sind neben geeigneten visuellen Metaphern insbesondere disziplinspezifische Verfahren der Daten-Manipulation und -Konfiguration zu bestimmen, die technisch als interaktive Manipulation von Visualisierungen umgesetzt werden können, um datenbasierte Forschungszugänge zu eröffnen und zu unterstützen.

Das skizzierte Spannungsverhältnis zwischen den allen Geisteswissenschaften gemeinsamen und den disziplinspezifischen Anforderungen an eine visuelles 'Expansionskonzept' hat Grinstein (2012) zur Formulierung einer 'grand challenge' motiviert:. Er fordert ein Visualisierungssystem, das auf disziplinspezifische Anforderungen reagiert und die in Hinblick auf die jeweilige Forschungsfrage wie die verfügbaren Daten optimale Visualisierungslösung automatisch generieren kann. Diese Vision mag zwar in der Tat 'grand' und unter dem Gesichtspunkt der Implementierbarkeit utopisch anmuten; als konzeptionelle Messlatte für das 3DH Projekt ist sie dennoch richtig. Denn nur Visualisierungslösungen, die den systematischen Zusammenhang zwischen den methodischen Anforderungen eines Forschungsvorhabens und den objektspezifischen Eigenschaften der in diesem Kontext erhobenen und generierten Daten konzeptionell reflektieren, haben zumindest eine theoretische Chance, die von Grinstein verlangten 'Passungen' automatisch zu ermitteln. [8]

Vorgehen und erste Ergebnisse

Das 3DH-Projekt erforscht den Phänomenbereich ‘Datenvisualisierung’ vor diesem Hintergrund unter drei systematischen Aspekten, nämlich

(1) einer Typologie hermeneutischer Routinen, Bedingungen und Zielsetzungen des begriffsorientierten (d.h. natürlich- bzw. fachsprachlich artikulierten) Interpretierens von Daten, die in ihrer für den geisteswissenschaftlichen Verstehensprozess kennzeichnenden Ausprägung zu definieren sind;

(2) einer Syntax grafischer Strategien, die – je nach Kontextbedingung und Prozessphase – die ‘bottom up’-definierten Grundlagen für ein erkenntnisproduktives visuelles ‘mapping’ der vorgenannten hermeneutischen Operationen auf die jeweils behandelten Primär- und Sekundärdatensets bereitstellen; und

(3) einer nach Designprinzipien geordneten Taxonomie konkreter Visualisierungstypen, die als ‘top-down’-Determinanten und epistemologische Paradigmen aufgefasst werden können. Die Designprinzipien werden ihrerseits nicht auf die Funktion der bloßen Steuerung visueller Datenrepräsentation am Ende eines geisteswissenschaftlichen Arbeitszyklus reduziert; sie sollen vielmehr als eigenständige, komplementäre Verfahren nicht-sprachlicher, bildgebundener Verstehensoperationen aufgefasst werden.

Die Bearbeitung der drei Aspekte soll neben der theoretischen Konzeptentwicklung auch zur Erarbeitung einer visuellen Grammatik für geisteswissenschaftliche Datenvisualisierung führen.

Im ersten Schritt haben wir eine Reihe exemplarischer Use Cases der DH-Forschung [9] betrachtet. In Anlehnung an Unsworths ‘scholarly primitives’ (Unsworth 2000) wurde untersucht, welche epistemologischen Prinzipien dabei für die Deutung und interaktive Bearbeitung von geisteswissenschaftlichen Daten wichtig waren. Diese Prinzipien können tabellarisch als Gegensatzpaare dargestellt werden:

Unreliability (inconsistency)	Reliability
Contradiction	Consent
Ambiguity	Definiteness
Uncertainty	Plausibility
Incompleteness (partial knowledge)	Comprehensiveness
Analogy	Identity
Probability	Factuality
Salience	Speculativeness

Tabelle 1: epistemologische Gegensatzpaare

Jedes dieser Gegensatzpaare markiert eine Dimension hermeneutischer Praxis, in der datenbasierte Erkenntnisprozesse in der Regel nicht auf normativ geregelte finite Auslegungen von Bedeutung und Wert, sondern auf kontextsensitive, skalierte dynamische Zuschreibungen von Informationsgehalt und Relevanz abzielen.

Als epistemologische Matrix bildet diese Tabelle zugleich die Grundlage für die Entwicklung einer ‘Grammar of Graphics’ in Anlehnung an Bertin (1983) und Wilkinson (2005). Wie von Satyanarayan et al. (2016) vorgeschlagen, müssen diese Ansätze allerdings um den Aspekt der Interaktivität erweitert werden. Graphische Merkmale sollen entsprechend durch sog. “Aktivatoren” visuell modalisierbar werden. [10] Der Grad an Unsicherheit einer spezifischen hermeneutischen Zuschreibung könnte z.B. visuell ausgedrückt werden, indem am Bildschirm *nachträglich* – also erst im Zuge der geisteswissenschaftlichen Dateninterpretation – die Transparenz einer Grafik interaktiv manipuliert und zugleich als Datenwert in der zugrundeliegenden Datentabelle erfasst wird.

Die so erweiterte visuelle Grammatik soll in eine Notation überführt werden, die möglichst allgemein verständlich, generisch und unabhängig von einer bestimmten Programmiersprache implementierbar sein muss; aufgrund der großen Verbreitung von XML in den Digital Humanities ist eine zusätzliche XML-Notation geplant. Daneben sollen für eine Reihe exemplarischer hermeneutischer Verstehens- und Interpretationsprozesse die systematischen Zusammenhängen zwischen Datenstrukturen und geeigneten Visualisierungsprinzipien erforscht und adäquate Vorschläge für eine (oder mehrere) Visualisierungen erarbeitet werden.

Die Implementierung der entwickelten Visualisierungen wird eine webbasierte

Browser-Anwendung sein, die kollaboratives Arbeiten ermöglicht und über ein Web Service Interface mit anderen Systemen verbunden werden kann. Die Spezifikation der Visualisierungen mit Hilfe einer von einer Grafik-Engine unabhängigen Grammatik erlaubt prinzipiell beliebige Ausgabeformate. Aufgrund der Interaktivität und Webfähigkeit ist zunächst SVG als Format geplant.

Ausblick

Auch wenn die weiteren Schritte zur Erarbeitung der visuellen Grammatik und der prototypischen Implementierung geisteswissenschaftlich funktionaler Visualisierungsansätze vorgezeichnet scheinen: Die Frage nach der methodischen Adäquatheit des Vorgehens bleibt für unser Vorhaben weiterhin virulent.

So stehen bei den epistemologischen Gegensatzpaaren in Tabelle 1 bislang logische Gegensätze des Typs A und non-A (z. B. Reliablity vs. Unreliability) und phänomenologische Gegensätze (z. B. Probability vs. Factuality) nebeneinander. Noch ist nicht geklärt, ob es sich hier um Kategorienfehler im analytischen Sinne handelt, oder ob nicht gerade dieses Nebeneinanderstehen kategorial unterschiedlicher Konzepte dem hermeneutischen Prozess gerecht wird. Welche Konsequenzen hätte es zum Beispiel für ein geisteswissenschaftliches Visualisierungskonzept, wenn sich strikt logische, binäre Modellierungen hermeneutischer Prozesse sogar als prinzipiell ungeeignet erweisen?

Unter diesem kritischen Vorbehalt erscheinen zum einen konkrete, etablierte visuelle Verfahren in einem neuen Licht. Kann zum Beispiel Shneidermans (1996) bekanntes *Overview, Zoom, Details on Demand*-Mantra für das geisteswissenschaftliche Arbeiten, das auf exemplarisches Sinnverstehen und nicht auf möglichst solide fundierte empirische Übersicht ausgerichtet ist, überhaupt Gültigkeit besitzen?

Erst das Nachdenken über die Erfordernisse eines geisteswissenschaftlichen Visualisierungs*konzepts* macht es zum anderen möglich, die epistemologische Funktion von Visualisierungen jenseits der bloßen Repräsentation von Datenpunkten auf einem Bildschirm zu begreifen. So gesehen steht die Praxis der Visualisierung als Expansion bzw. ‘Rückübersetzung’ und als Vermittlung zwischen Abstraktion und Phänomenologie in der philosophischen Tradition der *Aisthesis* - ein Aspekt, auf den Wilkinson (2005:1) verweist, wenn er feststellt: “Aesthetics, in the original Greek sense, offers principles for relating sensory attributes (color, shape, sound, etc.) to abstractions.”

Fußnoten

1. vgl. Goodings (2003:281): „Having reduced some aspect of the world to a form that can be processed according to rules, the output of the computation needs to be reintroduced into the world of meaningful, human action. [...] This involves translating the output into a familiar notational system and, in some cases, restoring more basic sensory modes of apprehension, as in the case of data visualization or the phenomenology of a thought experiment. [...] Instead of looking for cognitive capacities of the sort required by an algorithmic view of science as rule-based reasoning about an inherently digitizable world, we should investigate those cognitive capacities that enable practitioners from different cultures to exchange meanings and methods.“
2. Wir betrachten die Geisteswissenschaften nicht als Gegensatz zu den Naturwissenschaften oder Informationswissenschaften, sondern gehen vielmehr von einem Kontinuum aus, das sich zwischen den Polen Subjektivität/Einmaligkeit/Besonderheit und Objektivität/Reproduzierbarkeit/Allgemeingültigkeit entfaltet. Die Wechselwirkung zwischen Beobachtenden und Beobachtetem spielt nicht nur in der geisteswissenschaftlichen Hermeneutik oder den Sozialwissenschaften (z. B. in Kontext der Feldforschung), sondern auch in den eher als “objektiv” wahrgenommenen Naturwissenschaften eine Rolle, etwa in der Beobachtung von Heisenberg, dass sich die Wellenfunktion in der Quantenmechanik durch unsere Beobachtung ändert (Heisenberg 1959:37). Entsprechend sind die Ausführungen in diesem Beitrag potenziell für alle Wissenschaften bzw. Fragestellungen relevant, in denen hermeneutische oder analoge Prinzipien gelten.
3. Als Daten verstehen wir alle multimedialen bzw. intermedialen Primärdaten sowie das Gesamtspektrum an Meta- und Sekundärdaten und Verweisen, die auf diese referieren.
4. vgl. zu den speziellen Anforderungen für Visualisierungen in den Geisteswissenschaften z. B. Stone (2009) und Drucker (2011), sowie Windhager (2013) für einen Ansatz zur Umsetzung der Anforderungen.

5. Auch generalisierende Beiträge zur Visualisierung – wie etwa Ward et al. (2010) – klammern die disziplinär-methodischen Fragen aus, die mit der Visualisierung verbunden sind. In der Visualisierungscommunity setzt sich allerdings langsam das Bewusstsein um die Spezifik geisteswissenschaftlicher Daten durch. So stellen etwa die Organisator/innen des *Workshop on Visualization for the Digital Humanities* im Kontext der IEEE VIS 2016-Konferenz fest: „Despite the growing popularity of digital methods for research in the humanities, digital humanists are underserved by academics in visualization, and under-represented in visualization conferences". Diesen Mangel machen sie in disziplinären Unterschieden fest, die durch die interdisziplinäre Kommunikation über die Bedarfe der Geisteswissenschaften adressiert werden sollen [vgl. http://vis4dh.com/ , gesehen am 18.08.2016].
6. Das Projekt "3DH – dreidimensionale dynamische Datenvisualisierung und Exploration für Digital Humanities-Forschungen" wird in der ersten Projektphase (02/2016-01/2019) von der Behörde für Wissenschaft, Forschung und Gleichstellung gefördert. Für weitere Informationen vgl. www.threedh.net [gesehen am 18.08.2016].
7. Vgl. Sinclair et al. (2013:2) zur Rolle von interaktiven Visualisierungen: „Interactive visualizations [...] aim to explore available information, often as part of a process that is both sequential and iterative. That is, some steps come before others, but the researcher may revisit previous steps at a later stage and make different choices, informed by the outcomes produced in the interim."
8. Vgl. dazu auch Culy (2013), der Grinsteins 'grand challenge' einschätzt als „worth taking as a point of departure for the visualization of language and linguistic data."
9. Bei den Use Cases handelt es sich um insgesamt fünf laufende oder abgeschlossene DH-Forschungsprojekte der Projektmitglieder, die in der Gruppe intensiv im Hinblick auf die tatsächliche und mögliche Rolle von Visualisierungen im Forschungsprozess diskutiert wurden.
10. Diese graphischen Aktivatoren sind: tone (white to black/brightness), value (saturation), color (hue), transparency, texture, shape, orientation, position, size, resolution, blur, direction of motion, rate of movement, acceleration, rate of change, duration, form, surface, motion, sound (tone, volume, rhythm), voice, text.

Bibliographie

Bertin, Jacques (1983): *Semiology of Graphics*. University of Wisconsin Press.

Coles, Katharine (2016): *Show Ambiguity*. Workshop on Visualization for the Digital Humanities at IEEE VIS 2016. http://vis4dh.com/papers/Show%20Ambiguity%20Collaboration%20Anxiety%20and%20the%20PLeasures%20of%20Unknowing.pdf [letzter Zugriff 3. November 2016].

Culy, Chris (2013): „Tackling a grand challenge in the visualization of language and linguistic data", in: *DGfS 2013 Workshop on the Visualization of Linguistic Patterns*. http://ling.uni-konstanz.de/pages/home/hautli/LINGVIS/dgfs13_culy_abstract.pdf [letzter Zugriff 17. November 2016].

Drucker, Johanna (2011): „Humanities Approaches to Graphical Display", in: *DHQ: Digital Humanities Quarterly* 5 (1). http://digitalhumanities.org/dhq/vol/5/1/000091/000091.html [letzter Zugriff 17. November 2016].

Gadamer, Hans Georg (1972): *Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik* 3. Aufl. Tübingen: Mohr.

Gooding, David (2003): „Varying the Cognitive Span: Experimentation, Visualisation, and Computation", in: Radder, Hans (ed.): *The Philosophy of Scientific Experimentation*. Pittsburgh, PA: University of Pittsburgh Press 255–283.

Grinstein, Georges (2012): „New Grand Challenges in Information Visualization: New Theories, New Devices, and New Capabilities" in: *Keynote address at iV2012*.

Heisenberg, Werner (1959): *Physik und Philosophie*. Stuttgart: Hirzel.

Moretti, Franco (2013): „Operationalizing", in: *New Left Review* 84. https://newleftreview.org/II/84/franco-moretti-operationalizing [letzter Zugriff 30. November 2016].

Satyanarayan, Arvind / Dominik Moritz / Kanit Wongsuphasawat / Jeffrey Heer (2017): „Vega-Lite: A Grammar of Interactive Graphics", in: *IEEE Transactions on Visualization and Computer Graphics* 23 (1): 341–50 10.1109/TVCG.2016.2599030.

Shneiderman, Ben (1996): „The eyes have it: a task by data type taxonomy for information visualizations", in: *Proceedings of the IEEE Symposium on Visual Languages*. IEEE Computer Society Press, 336–43.

Sinclair, Stéfan / Ruecker, Stan / Radzikowska, Milena (2013): „Information Visualization for Humanities Scholars", in: Price,

Kenneth M. / Siemens, Ray (eds.) *Literary Studies in the Digital Age: An Evolving Anthology*. New York: Modern Language Associaton. https://dlsanthology.commons.mla.org/information-visualization-for-humanities-scholars/ [letzter Zugriff 17. November 2016].

Stone, Mareen (2009): „Information Visualization: Challenge for the Humanities“, in: *Working together or apart: Promoting the next generation of digital scholarship*. Washington, DC: Council on Library and Information Resources 43-56 https://www.clir.org/pubs/resources/promoting-digital-scholarship-ii-clir-neh/stone11_11.pdf [letzter Zugriff 17. November 2016].

Unsworth, John (2000): „Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?“, in: Symposium on *Humanities Computing: formal methods, experimental practice*. http://www.people.virginia.edu/~jmu2m/Kings.5-00/primitives.html [letzter Zugriff 17. November 2016].

Ward, Matthew / Grinstein, Georges / Keim, Daniel (2010): *Interactive data visualization: foundations, techniques, and applications*. Natick, Mass.: Peters.

Windhager, Florian (2013): „On Polycubism. Outlining a Dynamic Information Visualization Framework for the Humanities and Social Sciences“, in: Füllsack, Manfred (ed.): *Networking Networks: Origins, Applications, Experiments*. Wien; Berlin: Turia + Kant 28–63.

„Der Helmut Kohl unter den Brotaufstrichen“. Zur Extraktion vossianischer Antonomasien aus großen Zeitungskorpora

Jäschke, Robert
r.jaschke@sheffield.ac.uk
University of Sheffield

Strötgen, Jannik
jannik.stroetgen@mpi-inf.mpg.de
Max-Planck-Institut für Informatik, Saarbrücken

Krotova, Elena
kroelebor@gmail.com
Higher School of Economics, Moskau

Fischer, Frank
frafis@gmail.com
Higher School of Economics, Moskau

Einführung und Forschungslage

Wenn Peter Paul Rubens als »Tarantino des Barock« beschrieben wird (im *Tagesspiegel*, 2014) oder Alice Schwarzer als der »Erich Honecker des Feminismus« (in *Cicero*, 2014), dann handelt es sich um eine Vossianische Antonomasie. Diese Trope ist nach dem niederländischen Humanisten und Rhetoriklehrer Vossius benannt (und wird im Folgenden als ›Vossanto‹ abgekürzt, in Anlehnung an den Vorschlag von Fischer/Wälzholz 2014). Generell spricht man von Antonomasie, wenn eine bestimmte Eigenschaft einer Person für diese selbst steht (z. B. »der Leimener« für Boris Becker). Beim Spezialfall der Vossanto wird einer Person über die Nennung einer anderen (bekannteren, populäreren, berüchtigteren) Person als Referenzgröße eine bestimmte Eigenschaft zugeschrieben. Dabei sorgt ein »untypologisches, aktualisierendes Signal« (Lausberg 1960) für den Bedeutungstransfer (in den oben genannten Beispielen wären dies der Barock und der Feminismus). Anders ausgedrückt: Die Vossanto stellt über einen ›modifier‹ einen Zusammenhang zwischen ›source‹ und ›target‹ her (Bergien 2013). Entitäten können sowohl als ›source‹ als auch als ›target‹ auftreten, wie ebd. am Beispiel Obama demonstriert: bis 2011 trat er in Vossantos vor allem als ›target‹ auf, danach diente er immer mehr als ›source‹. Die ›source‹-Referenz wird im Fachdiskurs im Anschluss an Lakoff 1987 auch als ›paragon‹ bezeichnet (»a specific example that comes close to embodying the qualities of the ideal«, ebd.).

Der Begriff »Vossianische Antonomasie« wird international kaum verwendet, stattdessen wird etwa zwischen »Antonomasia1« und »Antonomasia2« unterschieden: »metonymic« vs. »metaphorical antonomasia« (Holmqvist/Płuciennik 2010). Innerhalb dieses Klassifikationsschemas wäre unsere Vossanto ein Spezialfall von »Antonomasia2«, nämlich wenn es um »comparisons with paragons from other spheres of culture« geht: »Lyotard

is a pope of postmodernism, Bush is no Demosthenes; and we can buy the Cadillac of vacuum cleaners.« (ebd.)

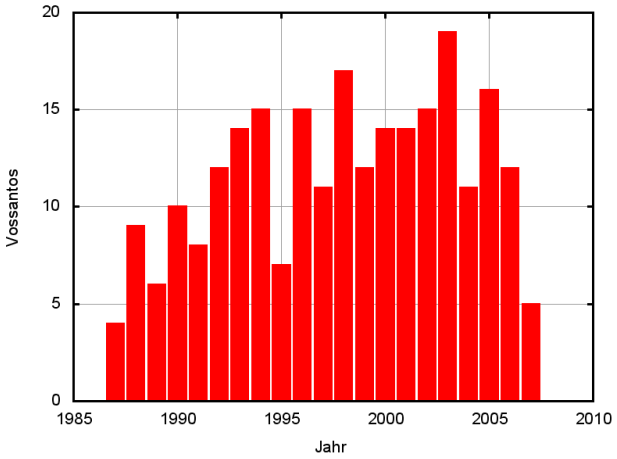
Dieses Stilmittel, dessen reger Gebrauch seit der Antike belegt ist, ist heute medial ubiquitär anzutreffen. Oft findet es sich schon in Überschriften, da es zugleich informativ und rätselhaft sein kann und zudem oft unterhaltsame Qualitäten bietet. Eine eigene größere Sammlung an Musterexemplaren (http://www.umblaetterer.de/datenzentrum/vossianische-antonomasien.html) gab den Ausschlag, dieses Phänomen systematisch zu erforschen, mit historischer Perspektive und auf Grundlage größerer englischer und deutscher Zeitungskorpora. Ziel dieser Arbeit ist eine erste methodisch-explorative Analyse des Phänomens Vossanto in der Tageszeitung *New York Times* (1987–2007) und der Wochenzeitung *Die Zeit* (1995–2011). Die Korpora wurden aufgrund ihrer Verfügbarkeit, Bedeutung und ihres Umfangs gewählt. Die Extraktion der Vossantos erfolgte jeweils korpusspezifisch, um den verschiedenen Formaten und Sprachen Rechnung zu tragen.

Englischsprachiges Korpus

Aus den XML-Daten des *New York Times*-Korpus (Sandhaus 2008) wurde für jeden der 1.854.726 Artikel der Volltext extrahiert. Anschließend wurde der Text mit Hilfe des NLTK (Bird/Loper/Klein 2009) in Sätze zerlegt und die Wörter jedes Satzes mit dem Part-of-Speech-Tagger des NLTK annotiert. Zusätzlich wurden Named Entities (Personen, Orte, Organisationen) mit dem NLTK-eigenen Named-Entity-Extraktor annotiert. Die so annotierten Sätze wurden mit einer Liste von Vossanto-typischen Mustern (in Form von regulären Ausdrücken) abgeglichen. Eine vereinfachte Darstellung eines solchen Musters ist beispielsweise:

> \((PERSON|ORGANIZATION|GPE) *\) (is|has) (often|sometimes)? (been)? (called)? the \(PERSON|ORGANIZATION|GPE) *\) (of|among|from) \((PERSON|ORGANIZATION|GPE) *\)

Die zu findenden drei Elemente einer Vossanto sind darin durch * gekennzeichnet. Passte ein Satz auf eines der Muster, so wurden diese drei Teile extrahiert und tabellarisch ausgegeben. Anschließend wurden die extrahierten 10.744 Kandidaten manuell überprüft. Nicht-Vossantos und Vossantos mit Städten und Firmen wurden entfernt (Treffer der Art »Algarve, the Riviera of Portugal« oder »Pepsi is the Nike of soft drinks« sind eine eigene Untersuchung wert) und der Fokus auf Vossantos gelegt, in denen Individuen (Personen, Tiere, fiktive Figuren) als ›source‹ oder ›target‹ dienen. 246 Vossantos blieben dabei übrig (Übersicht in unserem Arbeitsrepo, siehe Bibliografie), die sich wie folgt über das Korpus verteilen:



vossantos_nyt.png

Am häufigsten als ›source‹ verwendet wurden folgende Namen:

Anzahl	source
6	Michael Jordan
5	Michelangelo
4	Babe Ruth
3	Zelig
3	Rodney Dangerfield
3	Neil Young
3	Elvis
3	Don Quixote

Als Beispiel für Treffer seien diejenigen für Michael Jordan genannt:

- »Romario is the *Michael Jordan* of soccer and Bebeto is the Magic Johnson of soccer« (1994)
- »Bonfire, the *Michael Jordan* of dressage horses« (1998)
- »Brian Foster, the *Michael Jordan* of BMX racing« (1998)
- »The stunt biker Dave Mirra, the *Michael Jordan* of the dirt set« (2000)
- »Cynthia Cooper is the *Michael Jordan*, the Larry Bird, the Magic Johnson of this league« (2000)
- »McNabb has been called the *Michael Jordan* of the National Football League« (2001)

Trotz der zeitlichen Einschränkung des Korpus lassen sich bereits einige vielversprechende Beobachtungen anstellen und Thesen bilden: 1. Produktive Referenzgrößen einer Vossanto sind sowohl reale als auch fiktionale Figuren (Bsp. für letztere aus der obigen Liste: Woody Allens »Zelig«, Cervantes' »Don Quixote«). 2. Öffentliche Personen oder bekannte fiktionale Charaktere haben bestimmte Eigenschaften, die sie für die Verwendung als Referenzgröße einer Vossanto prädestinieren oder nicht (es bleibt etwa zu erforschen, warum gerade Michael Jordan und Michelangelo sich so gut eignen und nicht andere Sportler bzw. Künstler). 3. Es gibt historisch stabile Referenzgrößen, deren Bekanntheit vorausgesetzt werden kann (z. B. Michelangelo), und es gibt ephemere Referenzgrößen, die ab irgendeinem Zeitpunkt nicht mehr als Bezugspunkt taugen (für das benutzte zeitgenössische Korpus eher noch nicht relevant).

Deutschsprachiges Korpus

Das deutsche Datenset besteht aus einer Sammlung des Archivs der Wochenzeitung *Die Zeit* und enthält die Artikel aus den Jahren 1995 bis 2011. Insgesamt umfasst das Korpus 126.702 Dokumente.

Zunächst wurden die Volltexte (inklusive Überschriften) aller Dokumente extrahiert. Diese wurden dann mit Hilfe des Part-of-Speech-Taggers und Named-Entity-Recognition-Tools des Stanford CoreNLP Package verarbeitet. Für die Analyse deutschsprachiger Texte enthält Stanford CoreNLP speziell für das Deutsche trainierte Modelle (Faruqui und Pado 2010). Somit können alle Texte auf drei Ebenen untersucht werden: auf der Wortebene, der Part-of-Speech-Ebene sowie der Named-Entity-Ebene. Mithilfe von regulären Ausdrücken, die auf den verschiedenen Ebenen angewandt werden können, wurde dann nach Vossanto-Mustern gesucht. Im Gegensatz zur Verarbeitung des englischsprachigen Korpus wurde jedoch noch nicht versucht, auch das ›target‹ einer Vossanto zu extrahieren. Stattdessen wurden Muster entworfen, die das ›source‹-Objekt sowie das »aktualisierende Signal« matchen. Ausschlaggebend für diese Herangehensweise waren die in einem Testdurchlauf beobachtete hohe Anzahl an Vossantos ohne unmittelbaren Verweis auf das ›target‹ sowie eine große Vielfalt an möglichen Formulierungen, die auf die Relation zum ›target‹ hinweisen können. Mithilfe relativ strikter Regeln konnte die Anzahl an falschen Extraktionen im Rahmen gehalten werden. Ein vereinfachtes Beispiel für eine Extraktionsregel lautet etwa: »eine Art PERSON (der|des) (ADJECTIVE)? NOUN«.

Die Produktivität der beiden häufigsten Referenznamen des NYT-Korpus bestätigt sich im verwendeten deutschen Korpus, etwa wenn vom »Michael Jordan der analytischen Philosophie« die Rede ist (*Die Zeit* 44/1999) oder vom »bulgarischen Michelangelo« (*Die Zeit* 14/2001). Ansonsten scheint es sprachen- bzw. kulturspezifische Präferenzen zu geben. Die häufigsten ›sources‹ sind:

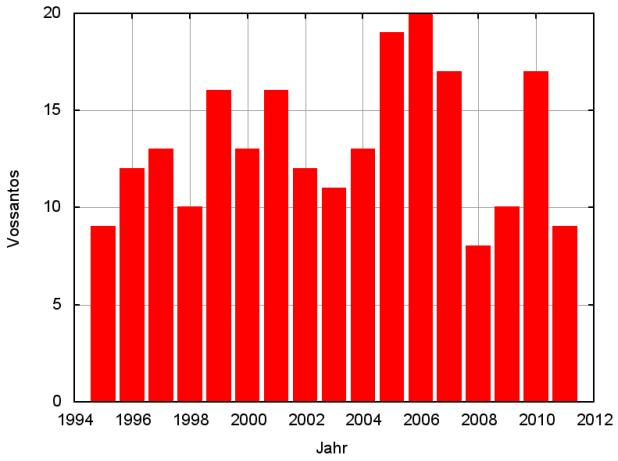
Anzahl	source
9	Robin Hood
6	Bill Gates
4	Franz Beckenbauer
3	Daniel Düsentrieb
3	Heinz Rühmann
3	James Dean
3	Jesus Christus
3	Norbert Blüm
3	Willy Brandt

Ähnlich wie im NYT-Korpus ist erkennbar, wie stark typisierend mythische bzw. fiktive Figuren sind (Robin Hood, Daniel Düsentrieb). Daneben zeigt sich, dass »Bill Gates«, der im NYT-Korpus nur zweimal als ›source‹ einer Vossanto vorkommt, im *Zeit*-Korpus sechs Mal als Referenz vertreten ist:

- »eine Art Bill Gates des Stolperns« (1998)
- »Der Bill Gates von Aurich« (2001)
- »der Bill Gates von Ostfriesland« (2001)
- »der Bill Gates von Aurich« (2002)
- »der britische Bill Gates« (2008)
- »der Bill Gates von Estland« (2010)

Die wiederholte Verwendung des »Bill Gates von Aurich« zeigt, wie stark ein ›target‹ mit einer ›source‹ verwachsen kann. (Paradebeispiel hierfür ist im Übrigen Vittorio Hösle, »der Boris Becker der Philosophie«, eine Bezeichnung, die es bis in den Wikipedia-Artikel zu Hösle geschafft hat.) Am Beispiel Bill Gates' lässt sich wie zuvor am Beispiel Obama demonstrieren, dass ein Name sowohl als ›target‹ als auch als ›source‹ vorkommen kann. Bevor Bill Gates selbst als Referenz verwendet wird, wird er in einem Artikel von 1995 noch durch eine andere Person beschrieben: »Bill Gates ist der Henry Ford des Computerzeitalters«.

Insgesamt wurden aus 1.456 Vossanto-Kandidaten 225 manuell als Vossantos markiert, die sich wie folgt über die im Korpus vorhandenen Jahre verteilen:



vossantos_zeit.png

Zu den fälschlich extrahierten Named Entities gehören »der Berliner Klaus Wowereit«, »der deutsche Michel« oder »der Anton aus Tirol«, stehende Wendungen, die grammatisch unseren definierten Vossanto-Mustern entsprechen.

Erkenntnisse und Ausblick

Die Vossanto ist als Stilmittel nur scheinbar einfach strukturiert, das Erstellen von Extraktionsregeln daher alles andere als trivial. Die vorliegenden Skripte weisen bekannte Lücken auf, die Qualität hängt v. a. von der Verlässlichkeit der benutzten NER-Tools und der Präzision der definierten Muster ab. Fehlende Goldannotationen für dieses Phänomen erschweren zudem eine Evaluierung. Die vorliegende Arbeit hat daher explorativen Charakter, die Optimierung von Precision und Recall lag noch nicht in deren Fokus, ist aber das nächste Ziel dieses Projekts.

Trotz der genannten Einschränkungen konnten durch diesen korpusbasierten Ansatz neue Erkenntnisse zur Vielgestaltigkeit des Phänomens ›Vossianische Antonomasie‹ gewonnen werden. So lassen sich zahlreiche Spezialfälle unterscheiden und systematisch untersuchen (vgl. auch Fischer/Wälzholz 2014), beispielhaft genannt seien:

- Tiere als ›target‹ (» *Sea Hero* is the Bobo Holloman of racing«, NYT, 1993; » *Bonfire*, the Michael Jordan of dressage horses«, NYT, 1998),
- Feminisierungen (Adele Schopenhauer, »eine Art *Donna Quichotta* des Weimarer Musenvereins«, *Die Zeit* 18/2002; »Tracey [Emin], die *Donna Giovanna* der britischen Gegenwartskunst«, *Die Zeit* 9/2006; »Kati Witt ist jetzt eine *Franziska Beckenbauer* der Münchner Olympiabewerbung.«, *Die Zeit* 39/2010),
- nicht individualisierbare ›sources‹: »the [God, King, Queen, Satan, Emperor, Oracle, Shogun, Czar, Sultan, Buddha] of«,
- mythologische und fiktive Figuren als ›sources‹: »the [Santa Claus, Midas, Godzilla, Pied Piper, Energizer Bunny, Jupiter, Icarus] of«,
- Personifizierungen, also der Einsatz individueller Personen/Figuren als ›source‹ für Firmen, Vereine, Bands oder Orte als ›target‹ (» *Sturm, Ruger* is the *Benedict Arnold* of the gun industry«, NYT, 1989; » *Aerosmith*, the *Dorian Gray* of rock bands«, NYT, 1993; »the *Hudson* has been the *John Barrymore* of rivers, noble in profile but a sorry wreck«, NYT, 1996; »the *National Collegiate Athletic Association*, the *Kenneth Starr* of sports«, NYT, 1998).

Zu letzteren Beispielen gehört nun endlich auch der titelgebende »Helmut Kohl unter den Brotaufstrichen« (*der Freitag* 35/2011).

Auch zur Distribution der Vossantos innerhalb der beiden Zeitungskorpora ließen sich belastbare Ergebnisse gewinnen. Demnach sind Vossantos besonders im Kultur- und Sport-Ressort beliebt (Vorkommen in der Sektion »Arts« der NYT: 78; in der Sektion »Sports«: 57; auf dem nächsten Rang mit großem Abstand »New York and Region«: 28 – im »Feuilleton + Literatur«-Ressort der Zeit: 76, »Politik«: 54, nächstrangig ist weit entfernt »Wirtschaft« mit 23 Vorkommen; »Sport« hat hier keine Treffer, denn die gedruckte *Zeit* hat kein dediziertes Sport-Ressort).

Bibliographie

Bergien, Angelika (2013): „Names as frames in current-day media discourse“, in: Felecan, Oliviu (ed.): *Name and Naming*. Proceedings of the second international conference on onomastics. Cluj-Napoca: Editura Mega 2013: 19–27.

Bird, Steven / Loper, Edward / Klein, Ewan (2009): *Natural Language Processing with Python*. O’Reilly Media Inc.

Faruqui, Manaal / Pado, Sebastian (2010): „Training and Evaluating a German Named Entity Recognizer with Semantic Generalization“, in: *Proceedings of Konvens 2010*.

Fischer, Frank / Wälzholz, Joseph (2014): „Jeder kann Napoleon sein: Vossianische Antonomasie: Eine Stilkunde“, in: *Frankfurter Allgemeine Sonntagszeitung* 51 (21. Dezember 2014): 34 http://www.umblaetterer.de/wp-content/uploads/2014/12/vossanto_fas.png .

Holmqvist Kenneth / Płuciennik Jarosław (2010): „Princess antonomasia and the truth: Two types of metonymic relations“, in: Burkhardt, Armin / Nerlich, Brigitte (eds.): *Tropical Truth(s): The Epistemology of Metaphor and Other Tropes*. Berlin/New York: De Gruyter 373–381 10.1515/9783110230215.

Lakoff, George (1987): *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: The University of Chicago Press.

Lausberg, Heinrich (1960): *Handbuch der literarischen Rhetorik. Eine Grundlegung der Literaturwissenschaft* 2. München: Hueber.

Sandhaus, Evan (2008): *The New York Times Annotated Corpus LDC2008T19*. DVD. Philadelphia: Linguistic Data Consortium.

Arbeitsrepositorium: https://github.com/weltliteratur/vossanto

Folien zum Vortrag: https://lehkost.github.io/slides/2017-bern/

Die Impactomatrix – ein interaktiver Katalog für Impactfaktoren und Erfolgskriterien für digitale Infrastrukuren in den Geisteswissenschaften

Thoden, Klaus
kthoden@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Deutschland

Wintergrün, Dirk
dwinter@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Deutschland

Stiller, Juliane
juliane.stiller@ibi.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Gnadt, Timo
gnadt@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland

Meiners, Hanna
meiners@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland

Einführung

Wissenschaftliche Großprojekte in den Geistes- und Kulturwissenschaften müssen sich damit auseinandersetzen, welchen Mehrwert sie für die wissenschaftliche Community schaffen, wie sie diesen sichtbar/messbar machen und wie sie die in sie investierten Mittel nutzbringend verwenden. Ausgehend davon war ein Forschungsziel in der ersten Förderphase von DARIAH-DE [1], dezidiert für die Geistes- und Kulturwissenschaften einsetzbare Erfolgskriterien und Impactfaktoren für digitale Tools und Infrastrukturkomponenten zu erheben. Dabei sollten nicht allein quantitative Merkmale wie Nutzungsstatistiken, sondern auch qualitative Merkmale wie beispielsweise Transparenz oder Nachhaltigkeit, berücksichtigt werden.

Die dabei zentralen Themen Erfolgsmessung, Impact und Evaluation sind bereits in einigen Publikationen – auch im Bereich der Digital Humanities – behandelt worden, beschränken sich jedoch in der Regel auf Nutzeranforderungen und -bedürfnisse für bestimmte Dienste und zu entwickelnde Tools (z.B. Brown u.a. 2006). Die Erfüllung dieser Anforderungen kann zwar als Erfolg gewertet werden, greift aber für eine umfassende Bewertung zu kurz. Genauso bieten Nutzerstudien einen Anhaltspunkt, wie Dienste und Tools genutzt und wo Verbesserungen angesetzt werden können. Beispielhaft soll hier die Nutzerstudie zu den Korpusplattformen, die bei der DHd 2016 vorgestellt wurde, genannt werden (Fandrych u.a. 2016).

Innerhalb des von der DFG geförderten Projektes "Erfolgskriterien für den Aufbau und nachhaltigen Betrieb von Virtuellen Forschungsumgebungen (DFG-VRE)" [2] wurde

ein generisches Set an Erfolgskriterien erstellt, welches an gegebene Projekte angepasst werden kann (Buddenbohm u.a. 2014) und nicht nur die Nutzerperspektive berücksichtigt, sondern auch interne Problematiken und Aspekte.

Zur tatsächlichen Messung von Veränderungen wurde im Rahmen der ersten Förderphase von DARIAH-DE, sowie in der *DARIAH-EU Working Group for impact factors and success criteria* [3] eine Übersicht entwickelt, die verschiedene Impact-Bereiche, diese Bereiche beeinflussende Faktoren sowie Kriterien zusammenträgt: Die *Impactomatrix* [4]. Ziel war neben der Bewertung der verschiedenen Kriterien und Faktoren unter Berücksichtigung verschiedener Stakeholder (WissenschaftlerInnen, BetreiberInnen, FörderInnen und EntwicklerInnen) auch ein modularer und erweiterbarer Aufbau.

Begriffe und Methodik

Den methodischen Untersuchungen ging eine extensive Literaturanalyse voraus, bei der Impact-Bereiche, und die diese beeinflussenden Kennzahlen und Faktoren extrahiert wurden – insgesamt konnten Begriffe aus 11 einschlägigen Quellen gezogen werden, die in Gnadt u.a. (2015) näher beschrieben sind. Basierend auf diesen Vorarbeiten wurden Erhebungen unter verschiedenen Stakeholdergruppen in Bezug auf digitale Tools und Infrastrukturdienste durchgeführt.

Innerhalb der geistes- und kulturwissenschaftlichen Community wurden zwei groß angelegte Online-Umfragen mit jeweils unterschiedlichen Zielgruppen vorgenommen: erstens 24 erfahrene, digital und in einem internationalen Kontext arbeitende FachwissenschaftlerInnen (Gnadt, Stiller & Höckendorff 2015) und zweitens 103 FachwissenschaftlerInnen, die hauptsächlich nicht digital arbeiten (Stiller u.a. 2015, Bulatovic u.a. 2016). Bei diesen Umfragen ging es vor allem um eine Einschätzung des Ist-Zustandes im Umgang mit digitalen Werkzeugen in der Forschung und der Nutzung von virtuellen Forschungsinfrastrukturen. Eine weitere Befragung zu Impactfaktoren und -kriterien fand unter den TeilnehmerInnen eines DINI-Workshops [5] sowie den DiensteanbieterInnen und DiensteentwicklerInnen in DARIAH-DE statt. Hierbei konnten die insgesamt 44 TeilnehmerInnen ihre Einschätzung der Wichtigkeit verschiedener Eigenschaften eines Tools abgeben, wie z.B. "Bedienbarkeit", "Funktionsumfang", "Dokumentation", "Einbeziehung von NutzerInnen", "curricularer Einsatz" und "Zahl an Referenzierungen". Zusätzlich wurden von einer Studentin im Rahmen ihrer Masterarbeit am Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin sechs Interviews mit VertreterInnen verschiedener Fachdisziplinen durchgeführt (Rose 2015). In den Interviews widmete sich die Autorin vor allem Fragen zur Einschätzung des Erfolgs von virtuellen Forschungsumgebungen und den eingesetzten Tools und Software in den jeweiligen Fachdisziplinen.

Auf Grundlage dieser Erhebungen und Studien wurde ein Katalog erarbeitet, der Impact-Bereiche, Faktoren und Kriterien zusammenfasst (Gnadt u.a. 2015). Diese Einteilung erfolgte auf der Basis der folgenden aus der Literatur abgeleiteten Definitionen von Impact, Erfolg, Kriterium und Faktor:

- *Impact* bezeichnet die Form, den Grad oder die Diversität einer Änderung eines Verhaltens oder Einstellung einer Gruppe
- *Erfolg* bezeichnet eine positive Resonanz auf eine Maßnahme oder ein Produkt, welche in ihrem Ausmaß messbar ist
- *Faktoren* beschreiben Eigenschaften oder Mittel zur Veränderung eines Zustands
- *Kriterien* beschreiben konkrete Merkmale zur Unterscheidung zwischen Zuständen

Abbildung 1 zeigt das Zusammenspiel von Impact-Bereichen, Faktoren, mit denen diese Bereiche beeinflusst und Kriterien, mit denen die Veränderungen gemessen werden können. Als Faktoren wurden auf der Basis der hergeleiteten Definitionen Eigenschaften, Mittel und Maßnahmen von Tools bzw. Forschungsinfrastrukturen klassifiziert, als Kriterien hingegen messbare Größen wie Kennzahlen, Indikatoren und Umfrageauswertungen. Der Erfolg von Tools und Forschungsinfrastrukturen wurde – ebenfalls auf der Basis der Literaturauswertungen – als Übereinstimmung von Nutzeranforderungen mit erreichtem Impact definiert.

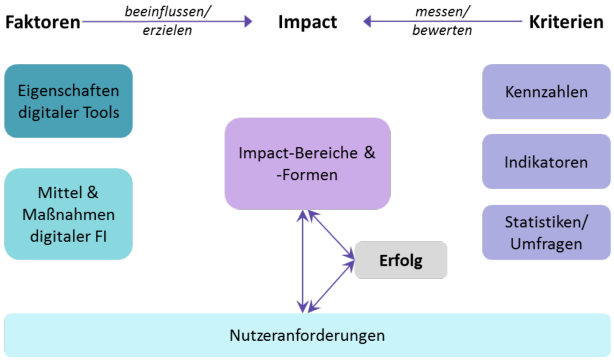


Abbildung : Zusammenspiel von Impact, Erfolg, Faktoren und Kriterien.

Die Impactomatrix

Aus den unterschiedlichen Erhebungen und der Literatur wurden Begrifflichkeiten für die Bereiche Impact, Kriterien und Faktoren gesammelt, ggf. übersetzt, zusammengefasst und in eine oder mehrere der drei Kategorien eingeordnet. Insgesamt wurden 101 relevante Begriffe extrahiert, von denen 21 als Impact-Bereiche identifiziert wurden. 67 Begriffe wurden als Faktoren eingestuft und 25 als Kriterien. [6] Bei einigen Begriffen gab es Mehrfachzuordnungen, da eine eindeutige Trennung nach Faktoren und Kriterien nicht immer möglich war.

Die Begriffe wurden außerdem ins Englische übertragen, mit dem Ziel, den Katalog einem größtmöglichen Publikum zugänglich zu machen. Auch die weitere Entwicklung der Impactomatrix wird auf Englisch erfolgen. Um diese gesammelten Daten nun für die Entwicklung, Anpassung und das Angebot von digitalen Diensten nutzen zu können, haben wir eine Übersicht in Form der Impactomatrix entwickelt und auf GitHub zur Verfügung gestellt. [7] Ausgehend von den 21 Impact-Bereichen können somit leicht die Faktoren ermittelt werden, die diese Bereiche beeinflussen. Zur positiven Veränderung von Impact in einem bestimmten Bereich sollten also diese Mittel eingesetzt bzw. diese Eigenschaften verbessert werden. Es können auch geeignete Kriterien bestimmt werden, mit denen Veränderungen im gegebenen Impact-Bereich gemessen werden können.

Die 21 Impact-Bereiche bzw. -Formen sind:

- Außenwirkung (External Impact)
- Bildung (Education)
- Datensicherheit/ Datenschutz (Data Security/ Safety)
- Dissemination (Dissemination)
- Effektivität (Effectivity)
- Effizienz (Efficiency)
- Förderperspektiven (Funding Perspective)
- Innovation (Innovation)
- Integration (Integration)
- Kohärenz (Coherence)
- Kollaboration (Collaboration)
- Kommunikation (Communication)
- Kompetenzvermittlung (Transfer of Expertise)
- Nachhaltigkeit (Sustainability)
- Nutzung (Usage)
- Publikationen (Publications)
- Relevanz (Relevance)
- Reputation (Reputation)
- Transparenz (Transparency)
- Wettbewerbsfähigkeit (Competitiveness)
- Wissenstransfer (Transfer of Knowledge)

Die Impactomatrix kann dazu verwendet werden, das Problembewusstsein für die Belange und Notwendigkeiten aller beteiligten Gruppen zu reflektieren und im Endeffekt erfolgreichere Angebote innerhalb einer Infrastruktur zu schaffen.

Zum Beispiel kann die Steigerung des Impacts in den Bereichen *Effizienz* und *Effektivität* erzielt werden, indem unter anderem Maßnahmen wie eine leichte Bedienbarkeit, die Einbettung in wissenschaftliche Workflows und die Bereitstellung von Hilfestellungen für die Nutzer umgesetzt werden. Ob diese Maßnahmen dann erfolgreich sind, um Effizienz und Effektivität zu steigern, kann in der Folge anhand verschiedener Indikatoren nachgewiesen werden. Solche Indikatoren sind beispielsweise die Intensität und der Umfang der Nutzung sowie das Ansehen und die Akzeptanz in der Community. Abbildung 2 zeigt einen Screenshot der Impactomatrix für den Bereich Nachhaltigkeit (Sustainability).

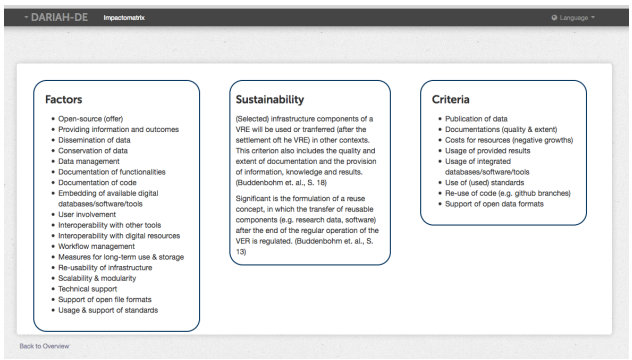


Abbildung : Screenshot der Impactomatrix.

Ausblick und Zusammenfassung

Die Impactomatrix bietet verschiedenen Stakeholdern einen Zugang zur Problematik der Erfolgsmessung von digitalen Tools und Diensten. Die Verzahnung der Impact-Bereiche mit ihren sie beeinflussenden Faktoren und Kriterien bietet eine einzigartige Möglichkeit, den Mehrwert von Entwicklungen in den Digital Humanities zu hinterfragen. Als Anwendungsbeispiel kann das Schreiben von Anträgen dienen, wenn es darum geht, auf den erhofften Impact des beantragten Projekts einzugehen sowie Maßnahmen zu bestimmen und festzulegen, die diesen Impact noch steigern können.

Hervorzuheben ist auch, dass in der Impactomatrix nicht nur die üblichen quantitativen Messzahlen in einer Liste zusammengetragen, sondern auch einen Beitrag zur qualitativen Bewertung von digitalen Diensten in den Geistes- und Kulturwissenschaften geleistet wurde: Ausgehend von dem Beispiel in Abbildung 2 ist ein Kriterium für die Nachhaltigkeit von Infrastrukturen die Unterstützung von offenen Datenformaten. Hier stellt sich die Frage, ob diese Kennzahl qualitativ (das meist genutzte offene Datenformat wird unterstützt) oder quantitativ (viele verschiedene offene Datenformate werden genutzt) gemessen werden sollte.

Die Impactomatrix wird ständig weiterentwickelt, und so wollen wir sukzessive eine engere Verzahnung der Faktoren mit den Kriterien (oder auch Kennzahlen) erzielen. Außerdem arbeiten wir an einer Ausdifferenzierung der Priorität verschiedener Impact-Bereiche für unterschiedliche Stakeholdergruppen. Da diese Weiterentwicklungen vornehmlich auf Feedback aus der Fachcommunity beruhen, laden wir mit diesem Beitrag auch dazu ein, den Katalog kennenzulernen, kritisch zu hinterfragen und Impulse für die Weiterentwicklung zu geben.

Fußnoten

1. https://de.dariah.eu/
2. https://www.sub.uni-goettingen.de/projekte-forschung/projektdetails/projekt/dfg-vre-1/
3. http://www.dariah.eu/activities/working-groups.html
4. Der Bewertungskatalog der zur Entwicklung der Impactomatrix geführt hat, wird im DARIAH-DE Report 1.3.3 (Gnadt u.a. 2015) ausführlich beschrieben.
5. https://dini.de/veranstaltungen/workshops/digitales-arbeiten-in-den-geisteswissenschaften-ermoeglichen
6. Das Ergebnis dieser Kategorisierung ist in Gnadt u.a. 2015 (Tabelle C.1 im Anhang C) zu sehen.
7. Quelltext auf https://github.com/DARIAH-DE/Impactomatrix , interaktive Version auf https://dariah-de.github.io/Impactomatrix/ .

Bibliographie

Buddenbohm, Stefan / Enke, Harry / Hofmann, Matthias / Klar, Jochen / Neuroth, Heike / Schwiegelshohn, Uwe (2014): *Erfolgskriterien für den Aufbau und nachhaltigen Betrieb Virtueller Forschungsumgebungen*. DARIAH-DE Working Papers 7 http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2014-5-4 [letzter Zugriff 29. November 2016].

Brown, Stephen / Ross, Robb / Gerrard, David / Greengrass, Mark / Bryson, Jared (2006): *RePAH: A User Requirements Analysis for Portals in the Arts and Humanities*. De Montfort University Leicester and The University of Sheffield http://repah.dmu.ac.uk/report/pdfs/RePAHReport-Complete.pdf [letzter Zugriff 29. November 2016].

Bulatovic, Natasa / Gnadt, Timo / Romanello, Matteo / Schmitt, Viola / Stiller, Juliane / Thoden, Klaus (2016): *Usability von DH-Tools und Services (R1.2.3)*. Göttingen: DARIAH-DE https://wiki.de.dariah.eu/download/attachments/14651583/AP1.2.3_Usability_von_DH-Tools_und-Services_final.pdf [letzter Zugriff 29. November 2016].

Fandrych, Christian / Frick, Elena / Hedeland, Hanna / Iliash, Anna / Jettka, Daniel / Meißner, Cordula / Schmidt, Thomas / Wallner, Franziska / Weigert, Kathrin (2016): „Wer bist du, Nutzer?“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 122–126 http://www.dhd2016.de/abstracts/vorträge-053.html [letzter Zugriff 29. November 2016].

Gnadt, Timo / Stiller, Juliane / Höckendorff, Mareike (2014): *Umfrage zu Erfolgskriterien (R1.3.1)*. Göttingen: DARIAH-DE https://wiki.de.dariah.eu/download/attachments/14651583/R%201.3.1%20-%20Erhebung%20einer%20Nutzerbefragung

%20zu%20Nutzererwartungen%20und%20-kriterien.pdf [letzter Zugriff 29. November 2016].

Gnadt, Timo / Stiller, Juliane / Thoden, Klaus / Schmitt, Viola (2015): *Finale Version. Erfolgskriterien (R1.3.3)*. Göttingen: DARIAH-DE https://wiki.de.dariah.eu/download/attachments/14651583/R133_Erfolgskriterien_Konsortium.pdf [letzter Zugriff 29. November 2016].

Rose, Corinna (2015): *Chancen und Grenzen der Abbildung fachspezifischer Forschungsprozesse durch eine virtuelle Forschungsumgebung in den Geisteswissenschaften*. Masterarbeit, Humboldt-Universität zu Berlin.

Stiller, Juliane / Thoden, Klaus / Leganovic, Oona / Heise, Christian / Höckendorff, Mareike / Gnadt, Timo (2015): *(R 1.2.1/M 7.6)*. Göttingen: DARIAH-DE https://wiki.de.dariah.eu/download/attachments/26150061/Report1.2.1-final.pdf [letzter Zugriff 29. November 2016].

Stiller, Juliane / Gnadt, Timo / Romanello, Matteo / Thoden, Klaus (2016): „Anforderungen ermitteln, Lösungen evaluieren und Erfolge messen – Begleitforschung in DARIAH-DE“, in: *Bibliothek Forschung und Praxis* 40 (2): 250–258. DOI:10.1515/bfp-2016-0025.

Digitale Modellierung literarischen Raums

Barth, Florian
florianbarth@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Viehhauser, Gabriel
viehhauser@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Problemstellung

Im Anschluss an den 1990 durch den Humangeographen Edward Soja ausgerufenen ‚Spatial Turn' (Soja 1990) haben sich zahlreiche kulturwissenschaftliche Forschungsarbeiten mit einer Beschreibung des Raums beschäftigt. In der Literaturwissenschaft fanden dabei u.a. kartografische Darstellungen große Resonanz: Franco Moretti etwa untersuchte in seinem „Atlas of the European Novel“ Orte der literarischen Produktion und Rezeption (Moretti 1998), Barbara Piattis Studie „Die Geographie der Literatur“ richtete den Fokus auf die Illustration einer konkreten literarisch thematisierten Gegend (die Zentralschweiz, vgl. Piatti 2008). Besondere Aufmerksamkeit wurde literarischen Karten auch im Kontext der Digital Humanities zu Teil, in denen geografische Informationssysteme (GIS) zum Einsatz kommen (typische Workflows beschreiben Gregory et. al. 2015)

Den meisten dieser Ansätze ist dabei gemein, dass sie für ihre Datengrundlage in erster Linie auf konkrete Nennungen von Ortsnamen (Toponymen) rekurrieren und weitere Ortsmarker weniger stark berücksichtigen. An der Konstitution literarischer Räume sind jedoch in der Regel auch komplexere Faktoren beteiligt, zu deren Beschreibung bereits erste narratologische Ansätze vorliegen (etwa von Kathrin Dennerlein [2009 und 2011] oder Gabriel Zoran [1984], vgl. auch die Überlegungen bei Piatti [2008]), die jedoch im Kontext der Digital Humanities bislang noch zu wenig Beachtung gefunden haben.

In unserem Beitrag möchten wir diese Ansätze aufgreifen, um das Instrumentarium der digitalen Textanalyse hinsichtlich der Kategorie des Raums zu schärfen und zu erweitern. Dazu scheinen uns insbesondere zwei Aspekte von Bedeutung: Zum einem die Unterscheidung von Raummarkierungen hinsichtlich ihrer Handlungsrelevanz (I), zum anderen die Ausweitung der Analyse auf räumliche Begriffe, die über bloße Namensnennungen hinausgehen (II). Für beide Problemfelder präsentieren wir erste Verfahren zur automatischen Auswertung und geben Ausblicke auf die Möglichkeiten einer vergleichenden Analyse.

I. Differenzierung von Räumen nach Handlungsrelevanz

Im Anschluss an Dennerleins Narratologie des Raumes hat sich insbesondere der Terminus der *räumlichen Gegebenheit* als Grundeinheit zur Bezeichnung von Ort und Raum durchgesetzt. Sind räumliche Gegebenheiten der Schauplatz eines konkreten Ereignisses, werden sie als *Ereignisregionen* spezifiziert. Diese haben als die zentralen handlungsrelevanten Räume besondere Bedeutung gegenüber *erwähnten räumlichen Gegebenheiten*, die bei nicht-situationsbezogener Thematisierung von Raum entstehen. In ähnlicher Weise unterscheiden Piatti (2008) sowie Piatti et. al. (2011) zwischen *Schauplatz* und *projizierten Orten*.

Am Beispiel von Jules Vernes „Reise um die Erde in 80 Tagen“ lässt sich die Wichtigkeit dieser Unterscheidung aufzeigen: So bietet z.B. Kapitel 14 eine Zugfahrt durch das Gangestal mit Aufenthalten in Allahabad und Benares sowie der Ankunft in Calcutta. Genannt werden im Text jedoch auch weitere Städte Indiens und das nächste Ziel Hongkong; eine Vielzahl der

extrahierten Ortsnamen bezieht sich somit nicht auf den Handlungsort des Kapitels.

Erst die kategoriale Trennung der Raummarkierungen in Ereignisregionen und erwähnte räumlichen Gegebenheiten ermöglicht die valide Rekonstruktion einer Reiseroute, die dann in einer GIS-Darstellung visualisiert werden kann (Abbildung 1).

Abbildung 1: Wegstrecke anhand zentraler Ereignisregionen in *Reise um die Erde in 80 Tagen* (QGIS-Grafik)

Eine automatische Unterscheidung dieser Kategorien muss daher das Fernziel computerunterstützter Raumuntersuchungen sein. Ansätze zu einer solchen Differenzierung suchen wir in der Anwendung von computerlinguistischen Methoden der Relationsextraktion, bei denen sich aus strukturellen Auffälligkeiten Regeln zur Klassifikation von Ereignisregionen und erwähnten räumlichen Gegebenheiten ableiten lassen. Hierzu zwei Beispiele:

1. Das 14. Kapitel der „Reise um die Erde in 80 Tagen“ beginnt mit dem in Abbildung 2 dargestellten Satz:

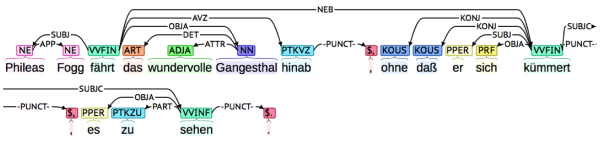


Abbildung 2: Dependenz-Pfad einer Ereignisregion (Malt Parser, Uni Tübingen)

Vor allem der Hauptsatz mit dem Pfad [Figur – SUBJ – Bewegungsverb - OBJ – Raumnomen] zwischen "Phileas Fogg" und "Gangesthal" kennzeichnet letzteres als Ereignisregion. Dabei stellt insbesondere die Verbkategorie ein Indiz für die Klassifikationsentscheidung dar: Statische Verben ("stehen", "sitzen") oder Verben der Bewegung ("gehen", "fahren") zeugen in spezifischen Satzstrukturen häufig von einer Ereignishaftigkeit im Gegensatz zu Verben der Kognition ("denken").

Zur Einteilung der Verben nutzen wir das von der Universität Tübingen entwickelte lexikalisch-semantische Netz *GermaNet* (Hamp/Feldweg 1997, Henrich/Hinrichs 2010), in dem eine Systematisierung vorliegt, die auf den Verbkategorien von Levin und dem Valenzwörterbuch von Schumacher basiert (Levin 1993, Schumacher 1986).

2. Findet ein erzähltes Ereignis innerhalb einer Bewegung im Raum statt, können mehrere Räume zu einem *Bewegungsbereich* zusammengefasst werden. Diese können bei Strukturen wie der folgenden leicht maschinell erkannt werden:

"Ich fuhr [...] *von Königstadt*, wo unser Hauptbüro war, *nach Gründerheim*, wo wir eine Nebenstelle hatten. *Dort* holte ich dringende Korrespondenz, Gelder und schwebende Fälle." (Böll, Heinrich: Über die Brücke)

Im zweiten Satz findet sich zudem eine Referenz auf das Antezedens "Gründerheim". Eigentlich werden derartige deiktische Adverbialausdrücke ("hier", "da" und "dort") bei der Koreferenz-Resolution nicht berücksichtigt. Deshalb planen wir eine Erweiterung der Trainingssets bestehender Koreferenz-Systeme hinsichtlich dieser Termini.

Die hier an zwei Beispielfällen dargestellten Regeln zur Unterscheidung narratologischer Raumeinheiten sollen in Zukunft kontinuierlich erweitert und zunächst so gestaltet werden, dass sie eine hohe Präzision erzielen. Anschließend können sie als Features für spätere maschinelle Lernverfahren verwendet werden.

II. Anreicherung der Raumbeschreibung

Netzwerkvisualisierung von Raumnomen

Eine kartografische Darstellung, wie sie in Abbildung 1 ersichtlich ist, bleibt auf realweltliche Ortsnamen beschränkt und lässt wesentliche Aspekte der Raumdarstellung außer Acht. Einen ersten Ansatz, die Vielfältigkeit der tatsächlichen Handlungsräume und ihrer Zusammenhänge abzubilden, bietet das Netzwerk in Abbildung 3. Hier werden für das angesprochene Kapitel 14 neben den Toponymen auch unspezifische Raumnomen als Knoten berücksichtigt und Verbindungen immer dann etabliert, wenn zwei Raumbegriffe gemeinsam in einem Satz auftreten.

Insbesondere landschaftliche (grün) und architektonische Raumnomen (grau) stellen relevante Klassen von Raummarkern dar, die neben den konkreten Ortsnamen zentrale Komponenten der literarischen Raumbeschreibung bilden. Als räumliche Gegebenheiten kommen aber auch bewegliche Objekte, in denen sich Figuren aufhalten

können, in Frage, wie die in der Grafik blau markierten Fahrzeuge.

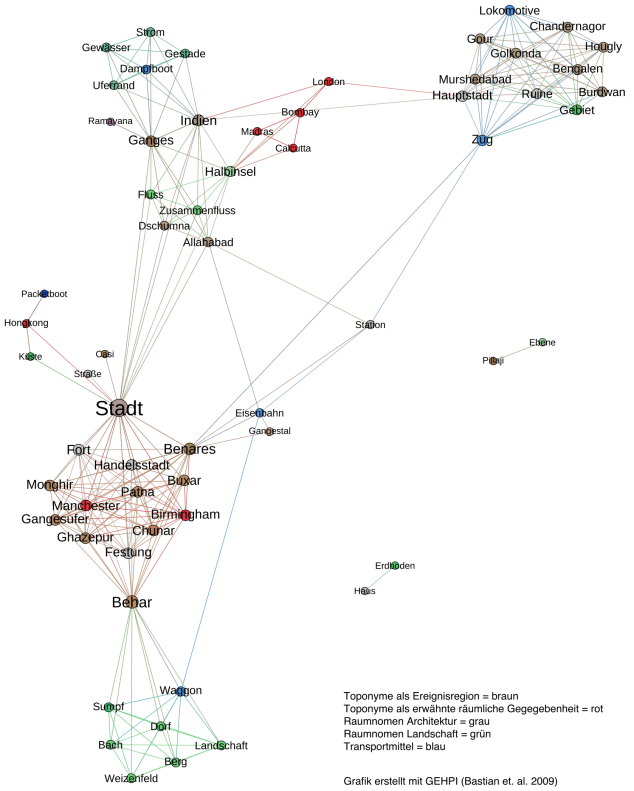


Abbildung 3: Netzwerk

Lexikon und Taxonomien für Raumbegriffe

Zur lexikalischen Erfassung von realweltlichen Toponymen greifen wir auf die Named-Entitiy-Recognition von *Weblicht* (2012) zurück, deren Ergebnisse wir mit dem Rückgriff auf die frei zugänglichen Datenbanken *GeoNames* (www.geonames.org) und *OpenStreetMap* (www.openstreetmap.org, Datendownload über www.geofabrik.de) zu verfeinern trachten. Aufgrund des Problems der möglichen Ambiguität von Ortsnamen (Leidner / Liebermann 2011, Gregory / Hardie 2011) ist jedoch eine manuelle Nachbearbeitung nötig. Listen von unspezifischen Raumnomen („Berg“, „Bach“, etc.) erstellen wir (ebenfalls semi-automatisch) auf der Basis von *GermaNet*.

Innerhalb dieses Lexikons planen wir zudem eine Einordung der Raumbegriffe in spezifische Taxonomien:

i) Vertikale Raum-Ort-Hierarchie

Narratologisch wird unter dem Begriff *Raum* ein umfassendes Gebiet in der erzählten Welt verstanden, welches ein Innen und Außen besitzt und wiederum lokalisierbare, punktuelle *Orte* beinhaltet (Dennerlein 2009). Diese Zuordnung erfolgt jedoch meist relational zur Erzählsituation: In Alfred Döblins Roman *Berlin Alexanderplatz* bildet die Stadt den Raum mit einzelnen Plätzen und Straßen als Orten. Unter einer geringfügigen Erweiterung des Erzählspektrums wäre Berlin aber potentiell nur ein Ort unter vielen im übergeordneten Raum Deutschland.

Statt einer festen Zuschreibung nähern wir uns dem Verhältnis von Ort und Raum über eine vertikale Taxonomie von räumlichen Gegebenheiten an, die von der Planetenebene bis zu jenen Objekten reicht, in denen sich unter Annahme faktualer Gesetzmäßigkeiten keine Figur mehr aufhalten kann. Im Sinne des *principles of minimum departure* (Ryan 1980) kann dabei so lange von einer nach realweltlichen Gesetzen eingerichteten Erzählwelt ausgegangen werden, bis deren bewusste Aufhebung innerhalb fiktionaler Texte in spezifischen Fällen eine gezielte Anpassung der Taxonomie erfordert (z.B. bei Aladins Flaschengeist in der Wunderlampe).

Abbildung 4 zeigt basierend auf den kapitelweise extrahierten Ereignisregionen in „Reise um die Erde in 80 Tagen“ die obersten taxonomischen Stufen: 1. Kontinent, 2. Land, 3. Stadt bzw. landschaftliche Region (inklusive Transportmittel, markiert in blau).

Europa					Afrika			Asien						
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Großbritannien					Ägypten		Ägypten (SK)	RM, SA, IO, Indien	Indien					
L	L	L	L	L (N)	S (N)	S	DM	DM, Aden, B	B, Zug	Zug, Callyan	Latanenwald bei Allahabad	P	P, A, Z, BE, Behar, C	C

Abbildung 4: Oberste Raumebenen der ersten 15 Kapitel in „Reise um die Erde in 80 Tagen“
Abkürzungen:(N) - Nebenfiguren
Europa: L - London
Afrika: S - Suez, SK - Sueskanal, DM - Dampfer Mongolia
Asien: RM - Rotes Meer, SA - Saudi Arabien, IO - Indischer Ozean, B - Bombay, A - Allahabad, P- Pillaji, BE - Benares, C - Calcutta

Während in den ersten beiden Ebenen dieses Beispiels ausschließlich Toponyme vorkommen, beinhaltet zumindest die dritte Stufe in der Nominalphrase „Latenenwald bei Allahabad“ ein unspezifisches Raumnomen. Weitere allgemeine Begriffe wären vor allem auf einer hier nicht dargestellten vierten Ebene zu finden (z.B. „Sumpf“, „Bach“, „Weizenfeld“ innerhalb der Landschaft „Behar“, vgl. Abb. 3).

Zur automatischen Erstellung einer solchen Hierarchie bieten sich bei Toponymen die in *GeoNames* vorhandenen Metadaten an, in denen bei jedem Stadt-Eintrag Informationen zu Land und Kontinent vorhanden sind. Bei unspezifischen Raumnomen eignet sich hingegen die hierarchische Struktur von

GermaNet. So stellt etwa der Begriff „Bach“ ein Hyponym zum übergeordneten Synset „Wasser/ Gewässer“ dar, letzteres besitzt wiederum die Hyperonyme „Land/Gegend/Gefilde“.

ii) Wortfelder

Wie in Abb. 3 ersichtlich, speisen sich Raumnomen zu großen Teilen aus den Wortfeldern Architektur und Landschaft. Die folgende Analyse basiert auf semiautomatisch erstellten Wortlisten, die auf der Basis von *GermaNet* durch die Auswertung der entsprechenden Synsets und Implikationen (Hyperonymie und Hyponymie) von zentralen Begriffen aus beiden Wortfeldern gewonnen wurden.

Makroperspektive

Das Potential digitaler korpusgestützter Raum-Analysen soll anhand des Vergleichs dreier ‚Berlin-Romane’ exemplarisch aufgezeigt werden. Dazu werden die Texte in jeweils zehn Segmente aufgeteilt und hinsichtlich der Frequenz spezifischer Raumbegriffe aus den Wortfeldern Architektur und Landschaft untersucht:

Dabei lassen sich deutlich höhere Anteile des architektonischen Vokabulars gegenüber dem Segmentmittelwerten eines Vergleichskorpus erkennen, das aus 451 im Textgrid-Repository enthaltenen Romanen besteht (Abbildung 5, oben).

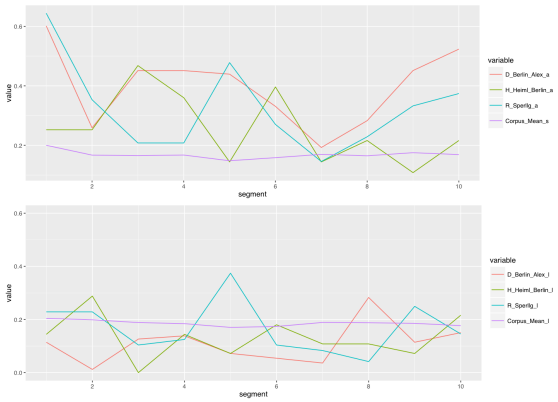


Abbildung 5: Architektonisches (oben) und landschaftliches (unten) Vokabular in *Berlin Alexanderplatz* (Alfred Döblin), *Heimliches Berlin* (Franz Hessel) und *Die Chronik der Sperlingsgasse* (Wilhelm Raabe) gegenüber dem mittleren Segmentverlauf im Korpus

Während die Verteilung des architektonischen Wortschatzes in Hesses „Heimliches Berlin“ nur temporäre Spitzen zeigt, sind die Segmentverteilungen von „Berlin Alexanderplatz“ und Wilhelm Raabes „Die Chronik der Sperlingsgasse“ gegenüber der mittleren Verteilung des Korpus signifikant verschieden. Dies wurde sowohl mit dem Wilcoxon-Rangsummentest (Annahme der Varianzhomogenität und Gleichverteilung zwischen den Sampleverteilungen) sowie dem Mood's Median-Test (keine Verteilungsannahme) überprüft (Abbildung 6).

Das landschaftliche Vokabular hingegen liegt bei den Berlin-Romanen tendenziell etwas unter dem Mittel des Korpus, allerdings sind die Abweichungen nur im Fall von „Berlin Alexanderplatz“ eindeutig signifikant (Abb. 5 unten, Abb. 6)

	Wilcoxon Architektur	Mood's Median Architektur	Wilcoxon Landschaft	Mood's Median Landschaft
Berlin Alexanderplatz	**0.0001224**	**0.001093334**	**0.001408**	**0.001093334**
Heimliches Berlin	0.07008	0.1788954	**0.0225**	0.1788954
Die Chronik der Sperlingsgasse	**0.0007523**	**0.001093334**	0.2406	0.6562818

Abbildung 6: p-Werte aus den Signifikanztests der Architektur- und Landschaftsverteilungen gegenüber den mittleren Segmenten im Korpus

Ungeachtet dieser Unterschiede sind die Zusammenhänge zwischen beiden Wortfeldern auffällig (Abbildung 7). Die Spearman-Korrelation zwischen architektonischen und landschaftlichen Begriffen bei „Berlin Alexanderplatz“ beträgt 0.5030488 und bei „Die Chronik der Sperlingsgasse“ sogar 0.7454545. So kann trotz abweichender Anteile der Wortfelder hinsichtlich ihrer Frequenz eine starke Verflechtung spezifischer Klassen von Räumen angenommen werden.

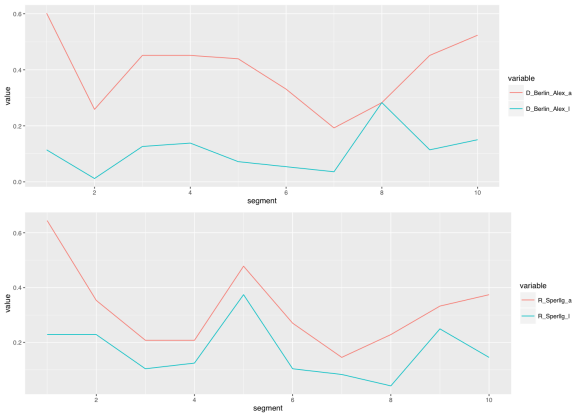


Abbildung 7: Korrelation zwischen den Wortfeldern Architektur und Landschaft in *Berlin Alexanderplatz* (oben) und *Die Chronik der Sperlingsgasse* (unten)

Ausblick

Die vorgestellten Ansätze verstehen sich als Anregung für die Entwicklung eines differenzierten Instrumentariums

der digitalen Raumanalyse, das in Zukunft weiter ausgebaut werden soll und die Grundlage für die Behandlung weiterführender literaturwissenschaftlicher Fragestellungen bildet, die etwa Aspekte der Semantisierung von Räumen (Lotman 1972), des raumzeitlichen Entwurfs von Erzählwelten (Bachtin 1989) und der Bedeutung von Raumkonstellationen für die Gattungspoetik beinhalten (vgl. zusammenfassend Nünning 2009).

Bibliographie

Bachtin, Michail Michailowitsch (1989): *Formen der Zeit im Roman. Untersuchungen zur historischen Poetik*. Ed. von Kowalski, Edward / Wegner, Michael. Frankfurt am Main: Fischer.

Bastian Mathieu / Heymann Sebastian / Jacomy Mathieu (2009): „Gephi. An open source software for exploring and manipulating networks“, in: *International AAAI Conference on Weblogs and Social Media*.

Dennerlein, Katrin (2009): *Narratologie des Raumes*. Berlin: de Gruyter.

Dennerlein, Katrin (2011): „Raum“, in: Matías Martínez (ed.): *Handbuch Erzähllliteratur: Theorie, Analyse, Geschichte*. Stuttgart / Weimar: Metzler 158–165.

Gregory, Ian / Hardie, Andrew (2011): „Visual GISting: bringing together corpus linguistics and Geographical Information Systems“, in; *LLC* 26: 297–314.

Gregory, Ian / Cooper, David / Hardie, Andrew / Rayson, Paul (2015): „Spatializing and Analyzing Digital Texts. Corpora, GIS, and Places“, in: David Bodenhamer / John Corrigan / Trevor Harris: *Deep Maps and Spatial Narratives*. Bloomington: Indiana University Press 150–178.

Hamp, Birgit / Feldweg, Helmut (1997): „GermaNet - a Lexical-Semantic Net for German“, in: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid.

Henrich, Verena / Hinrichs Erhard (2010): „GernEdiT - The GermaNet Editing Tool“, in: *Proceedings of LREC 2010* 2228–2235.

Leidner, Jochen / Lieberman, Michael (2011): „Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language“, in: *SIGSPATIAL Special* 3: 5–11.

Levin, Beth (1993): *English Verb Classes and Alternations*. University of Chicago Press.

Lotman, Juri (1972): *Die Struktur literarischer Texte*. München: Fink.

Moretti, Franco (1998): *Atlas of the European novel. 1800-1900*. London / New York: Verso.

Nünning, Ansgar (2009): „Formen und Funktionen literarischer Raumdarstellung: Grundlagen, Ansätze, narratologische Kategorien und neue Perspektiven“, in: Wolfgang Hallet / Birgit Neumann (eds.): *Raum und Bewegung in der Literatur: Die Literaturwissenschaften und der Spatial Turn*. Bielefeld: Transcript 33–52.

Piatti, Barbara (2008): *Die Geographie der Literatur. Schauplätze, Handlungsräume, Raumphantasien*. Göttingen: Wallstein.

Piatti, Barbara / Reuschel, Anne-Kathrin / Hurni, Lorenz (2011): „A Literary Atlas of Europe – Analysing the Geography of Fiction with an Interactive Mapping and Visualisation System“, in: *Proceedings of the 25th International Cartographic Conference*. Paris.

Ryan, Marie Laure (1980): „Fiction, Non-Factuals, and Minimal Departure“, in: *Poetics* 8: 403–422.

Schumacher, Helmut (1986): *Verben in Feldern: Valenzwörterbuch zur Syntax und Semantik deutscher Verben*. Berlin / New York: de Gruyter Verlag,

Soja, Edward (1990): *Postmodern Geographies: The Reassertion of Space in Critical Social Theory*. London / New York: Verso.

WebLicht (2012): CLARIN-D/SfS-Uni. Tübingen 2012. *WebLicht: Web-Based Linguistic Chaining Tool*. https://weblicht.sfs.uni-tuebingen.de/ [letzter Zugriff 1. Dezember 2016]

Zoran, Gabriel (1984): „Towards a theory of space in narrative“, in: *Poetics Today* 5: 309–335.

Digitale Transformationen. Zum Einfluss der Digitalisierung auf die musikwissenschaftliche Editionsarbeit

Meise, Bianca
bianca.meise@upb.de
Universität Paderborn, Deutschland

Meister, Dorothee
dm@upb.de
Universität Paderborn, Deutschland

Zusammenfassung: Digitale Daten stellen den zentralen Forschungsfokus der Digital Humanities dar. Fragen der Modellierung, Repräsentations-, Analyse- und Annotationsmöglichkeiten sind dabei wichtige Forschungsdimensionen, ebenso wie etwa die Weiterverarbeitung und Nachnutzbarkeit. Die digitalen Daten sowie die beschrieben Prozeduren werden jedoch auch von EditorInnen bearbeitet und wirken sich auf deren wissenschaftliche Arbeit aus. In diesem Beitrag wird aus qualitativ empirischer Sicht die Perspektive der EditorInnen als besondere Nutzer- und Produzentengruppen im Prozess der Digitalisierung von Musikeditionen vorgestellt. Dabei gilt es weder WissenschaftlerInnen noch Daten singulär zu betrachten, sondern im Akt der Bearbeitung, Analyse, Repräsentation und Annotation eine besondere Perspektive in der Auseinandersetzung von Medien, Materialien und Subjekten zu erschließen und zu reflektieren. In diesem Sinne werden in diesem Abstract zuerst theoretische Verortungen für die Relevanz des Nutzers diskutiert. Darauf aufbauend werden die methodischen Grundlagen der Interviewstudie vorgestellt, um anschließend einen Ausblick auf die Ergebnisse zu geben, der im Vortrag vertieft wird. Dabei stehen die Veränderungen des wissenschaftlichen Arbeitsprozesses von analog zu digital im Vordergrund. Darauf aufbauend stehen die Chancen und Herausforderungen dieses Paradigmenwechsels im Zentrum des Interesses, die sicherlich nicht nur für die Arbeitskontexte der digitalen Musikeditionen zutreffen. Abschließend werden Kristallisationspunkte und Konsequenzen zukünftiger Fragestellungen hinsichtlich der Digitalisierung von Musikeditionen, Veränderungen von Arbeitsstrukturen sowie der Bildungs- und Wissensarbeit resultierend aus diesen Ergebnissen thematisiert.

Einleitung

Digitale Musikeditionen bieten potentiell vielerlei Optionen für EditorInnen, WissenschaftlerInnen und geneigte RezipientInnen (vgl. etwa Veit 2010). So beschleunigt und ordnet etwa die Verfügbarkeit digitaler Quellen den editorischen Prozess, die Ansichten des Digitalen ermöglichen größtmögliche Transparenz und Nachvollzieharkeit. Umbestritten sind also die Errungenschaften, die mit den digitalen Musikeditionen einhergehen eine Bereicherung der wissenschaftlichen Arbeit (vgl. ebd.). Vieles was im Kontext der Digital Humanities diskutiert wird, bezieht sich auf die digitale Repräsentation oder aber Transformation der kulturellen Artefakte, deren weitergehenden Analyse bzw. Prozessierbarkeit sowie deren Nachnutzbarkeit. Diese Ebenen werden aus dem Fokus auf die Daten heraus diskutiert. Was hingegen weniger betrachtet wird ist die Forschung im Digitalen als Wissensgenerierungsprozess: Was bedeutet es, sich digitale Techniken anzueignen, digital Quellen zu bearbeiten, zu repräsentieren, zu analysieren? Die Auseinandersetzung mit diesen Praktiken der Aneignung ermöglicht es ein tieferes Verständnis für die Optionen der digitalen Daten und deren wissenschaftliche Relevanz im Arbeitsprozess zu eröffnen. Ebenso lassen sich die zuvor skizzierten Forschungsthemen von Repräsentation, Transformation, Nachnutzbarkeit etc. ebenso aus der Sicht der Subjekte erschließen.

Theoretische Rahmung

Neben Fragen der Daten werden somit zunehmend die Arbeitsprozesse interessant, die durch die Digitalisierung der wissenschaftlichen Arbeit beeinflusst werden. An dieser Stelle bietet sich die seltene Gelegenheit die Veränderungsprozesse dieses medialen Paradigmenwechsel und dessen Einfluss auf Forschung und Wissenschaft zu beobachten und zu begleiten. Damit gilt es die NutzerInnen in den Blick zu nehmen (vgl. auch Stone 1982, Edwards 2012, Warwick 2012; Brockman 2001) und vom forensic zum formal layer (vgl. Kirschenbaum 2008) zu wechseln. Aber auch Kirschenbaums formal layer bringt nicht ganz zum Ausdruck, was Drucker (2013) mit der performativen Ebene von Materialität als Nutzungsakt beschreibt: Handeln, der Umgang von NutzerInnen mit kulturellen, auch immateriellen Artefakten, prägen die Wahrnehmung, Beurteilung und die kulturelle Bedeutung dieser Artefakte. Um die Bedeutung von Medien, konkreter von musikeditorischen Ergebnissen unter digitalen Bedingungen, erschließen zu können, ist es notwendig, die vielschichtigen Auseinandersetzungsprozesse der Nutzer mit der Software bzw. der Auszeichnungssprachen und Metadaten zu erforschen. Damit verbunden ist die sogenannte radikale Kontextualisierung in den Cultural Studies, bei der davon ausgegangen wird, dass »Objekt und Subjekt, Medientechnologie und Kontext« (vgl. Winter 2010): sich stetig beeinflussen und miteinander verwoben sind. Erst in der Analyse dieser komplexen Verbindungen kann letztlich das Phänomen konturiert und erforscht werden. Medientechnologien und ihre Nutzer gehen demnach in zahlreichen Auseinandersetzungsprozessen eine Allianz ein, die in dieser Perspektive eine besondere

Qualität hervorbringt. Einen Schritt weiter geht Rainer Winter, indem er mit Rekurs auf Heidegger darauf hinweist, dass Medien nicht nur technische Artefakte sind, sondern gerade in ihrer Einbettung in soziale und kulturelle Prozesse, Optionen und Zugänge zur Welt umgestalten (vgl. ebd.). In dieser Hinsicht gilt es weitergehend Wissensgenerierungsprozesse in den Blick zu nehmen. In diesem Beitrag stehen die EditorInnen als besondere Nutzergruppe im Zentrum des Interesses. Diese arbeiten an der Schnittstelle vom computer und cultural layer (Manovich 2001). Sie arbeiten mit Metadaten und Auszeichnungssprachen und müssen somit die Logiken des Prozessierens des Computers verstehen, gleichzeitig arbeiten sie mit den Transformationen an der Oberfläche, lassen sich Teile oder Überblick bestimmte Werkaspekte anzeigen, um editorische Entscheidungen zu treffen und bilden damit einen ganz versierten Nutzer- und Produzententypus ab.

Methode

Im Projekt wurde auf die Prinzipien der qualitativen, empirischen Sozialforschung zurückgegriffen, um diesen Prozess möglichst offen und gegenstandsangemessen erforschen zu können (vgl. etwa Flick et al. 2000). Qualitative Sozialforschung birgt zunächst den Vorteil in einem unbekannten Forschungsfeld einsetzbar zu sein. Ergänzend zur quantitativen Forschung wird somit im Projekt ein hypothesengenerierendes und exploratives Verfahren verfolgt. Um die Sicht der EditorInnen in der Annäherung zu erschließen, können die Befragten durch unterschiedliche Methoden beforscht werden (Beobachtung, unterschiedliche Arten von Interviews, Einzel- bzw. Gruppeninterviews etc. vgl. ebd.). Die Auswahl der Erhebungsmethode erfolgt dem Forschungsphänomen entsprechend angemessen. Erhebt also die quantitative Forschung Daten zu vielen Nutzern, um Überblicke zu generieren und Themenfelder zu identifizieren ist die qualitative Forschung mit diesen Erhebungsmethoden dazu in der Lage in die Tiefe zu gehen und bspw. Bedeutungskontexte, implizites Wissen und unbewusste Routinen zu eruieren. Hierbei werden nicht nur Stichworte, sondern Kontextinformationen aus der Sicht der Subjekte aus den Daten herauszuarbeiten (vgl. Flick 2002). Die Erforschung impliziten Wissens, von Arbeitsroutinen, Expertisen und Gewissheiten lässt sich dabei kaum über einzelne direkte Fragen realisieren. Um solchen Phänomenen auszuspüren ist zunächst der Gesamtkontext wichtig. In diesem Sinne nutzt die qualitative Forschung verschiedene Formen von Befragungstechniken, um unterschiedliche Arten von Narrationen zu erhalten, die im Anschluss verschriftlicht und analysiert werden können. Im Auswertungsprozess wird dann über Interpretationen der Gesamtkontext erarbeitet und erschlossen. Dies geschieht indem einzelne Wissensbestände mit anderen Aussagen verbunden werden, die im Gesamtkontext Einblicke in das Zusammenwirken expliziter und impliziter Wissensbestände erlauben. Als Erhebungsform für diese qualitative Studie wurde das teilstandardisierte, narrative Interview gewählt, das zwar einem Leitfaden folgt, in der Interviewsituation allerdings größtmögliche Spielräume hinsichtlich der Frageformulierungen, Nachfragestrategien und der Reihenfolge der Fragen zulässt (vgl. Keuneke 2000) und der Narration der EditorInnen viel Raum gestattet. Der Leitfaden fokussierte besonders Regelstrukturen der Handlungs- und Nutzungsweisen, indem die Befragten nach bestimmten Nutzungssituationen und den damit verbundenen Bedeutungen und Relevanzen befragt wurden. Diese Nutzungssituationen und Bedeutungen wurden allen interviewten Personen gleichermaßen vorgegeben. Die von den Befragten formulierten Antworten konnten anschließend aufeinander bezogen werden.
Zur Auswahl der Interviewpartner/innen wurden im Sinne des Theoretical Sampling (vgl. Przyborski/ Wohlrab-Sahr 2009) Expertinnen und Experten, die mit Edirom arbeiten befragt. Die Software Edirom erlaubt es Faksimiles, Digitalisate und digitale Daten von Notentexten oder anderen Quellen einzuarbeiten, zu speichern, zu organisieren, zu kollationieren, zu annotieren und zu analysieren. Dabei handelt es sich nicht um eine Forschungsoberfläche, die Voraussetzungslos für die EditorInnen ist, sondern hier sind Auseinandersetzungen und Erfahrungen mit XML, TEI und MEI erforderlich. Entscheidend war, dass sowohl weibliche als männliche Nutzer/innen befragt wurden. Insgesamt wurden acht Interviews mit sechs Editorinnen und zwei Editoren geführt. Diese dauerten zwischen 90 Min. und 180 Min.
Die erhobenen qualitativen Daten wurden durch eine Variante des Kodierens [1] nach Strauss und Corbin, wie es Przyborski und Wohlrab-Sahr vorschlagen, ausgewertet (vgl. ebd.). Die Auswertung wurde zunächst nicht mittels Auswertungsprogrammen strukturiert, sondern durch Textverarbeitungsprogramme. So konnten die sich herausbildenden Phänomene einer exemplarischen, interdisziplinären Sichtung unterzogen werden. In der Synthese der exemplarischen Auswertung konnte das

selektive Kodieren vorangetrieben werden, woraus eine Phänomen- und Kategorienliste resultierte. Diese liefert einerseits konkrete Hinweise für Optimierungen der Software, aber auch Kontextinformationen zu den Arbeitsbedingungen, Routinen und Erfahrungen der EditorInnen. Darüber hinaus liefern die Interviews sehr gute Einblicke in die Änderungsprozesse der Wissensarbeit, des Wissensmanagements als auch des erarbeiteten Wissens als solches.

Ausblick auf die Ergebnisse

Wie die empirischen Daten belegen, ist der Wechsel der Arbeit und der beforschten Gegenstände von analog zu digital nicht nur eine technische Änderung, vielmehr gehen damit auch editorische, rechtliche, organisatorische, soziale und nicht zuletzt auch bildungswissenschaftliche Prozesse einher, die betrachtet, reflektiert und weiterentwickelt werden müssen. Sie verändern die wissenschaftliche Arbeitsorganisation, die editorische Tätigkeit und nicht zuletzt die Sicht auf Editionen und die damit verbundenen Erkenntnissen selbst. EditorInnen recherchieren bei ihrer analogen Forschungsroutinen zunächst Quellen, analysieren diese und wählen Haupt- und Nebenquellen aus, um die weitere Editionsarbeit zu gestalten. Im Anschluss daran wurden diese Quellen stetig miteinander verglichen. Dazu musste sehr viel Quellenmaterial physisch verwaltet werden, um die einzelnen Änderungen in der jeweiligen Quelle mit anderen vergleichen und analysieren zu können. Die Arbeit an den digitalen Editionen ist indes ein Konglomerat aus analogen und digitalen Techniken. Zuerst werden die Quellen ebenso recherchiert, analysiert und ausgewertet und ausgewählt. Die ausgewählten Quellen werden von den Hilfskräften im Anschluss digitalisiert, vertaktet und die Konkordanzen festgelegt. Es findet also eine Arbeitsteilung statt, da die Vertaktung delegiert wird. Die Interviews belegen die hohen Vorteile und Freiheitsgrade der digitalen Editionen, den Rezipienten solcher Editionen kann nun erstmals das gesamte Quellenmaterial zur Verfügung gestellt werden. Diese können nun editorische Entscheidungen transparent nachvollziehen und eine eigene Meinung dazu entwickeln. Damit einhergehend sind aber auch ein zunehmendes Maß an Komplexitätssteigerung und wachsenden Aufgaben zu verzeichnen. Bei Printeditionen steht die editorische Tätigkeit im Fokus. Der Wechsel zu digitalen Editionen bedeutet für die Editoren einen weiteren Komplexitätsschub: Nicht nur die musikwissenschaftliche Expertise ist gefragt, sondern auch Kenntnisse verschiedenster Auszeichnungssprachen, wie XML, TEI und MEI. Durch die Arbeitsteilung muss darüber hinaus den Hilfskräften Wissen für die Vertaktung vermittelt, diese angeleitet und kontrolliert werden. Zudem nutzen die EditorInnen notwendigerweise mehr Programme. Um nur einige zu nennen sind dies: Sibelius, Score, Finale, QuarkX, Indesign, OxygenXML, Lillypond, Word, Filemaker, oder aber Verovio. Wie die aufgeführten Notensatzprogramme verdeutlichen, sind EditorInnen nun teilweise auch mit Aufgaben beschäftigt, die vorher von Verlagen erledigt wurden. Durch diese Tätigkeit kann auch das Rechtemanagement von Originalquellen zu einem weiteren Aufgabengebiet werden. Die Ergebnisse abstrahierend betrachtet sind im neuen Medium neue Forschungsfragen entstanden und bilden sich täglich neu aus: Wo sind die Anfangs- und Endpunkte von Editionen, welche Nachnutzbarkeit kann gewährleistet werden, wie kann die Praxis von dem Wissen profitieren und dieses einsehen, wo ist Wissen gesichert erschlossen? Zudem gibt es kaum verbindliche Standards im digitalen Editionsprozess, was nun, bei steigender Editionszahl und entsprechender Annotationsmenge immer offensichtlicher und wichtiger wird. Ein systematischer Wissensaufbau informatischer Grundkenntnisse wird ebenso implizit evident, um die Potenziale der digitalen Repräsentations- und Verarbeitungsoptionen besser erschließen zu können.

Fußnoten

1. Das Kodieren bezeichnet im Gegensatz zu Semantiken aus der Informatik im Kontext der qualitativen Forschung eine Auswertungstechnik.

Bibliographie

Brockman, William S. / Neumann, Laura / Palmer, Carole L. / Tidline, Tonyia J. (2001): *Scholarly Work in the Humanities and the Evolving Information Environment*. portal: Libraries and the Academy (Vol. 3). Digital Library Federation. 10.1353/pla.2003.0012 .

Drucker, Johanna (2013): „Performative Materiality and Theoretical Approaches to Interface“, in: *DHQ: Digital Humanities Quartely* 7 (1). http://www.Digitalhumanities.org/dhq/vol/7/1/000143/000143.html .

Edwards, Charlie (2012): „The Digital Humanities and Its Users“, in: Gold, Matthew K. (ed.): *Debates in the Humanities*. http://dhdebates.gc.cuny.edu/debates/text/31 [letzter Zugriff 3. September 2015].

Flick, Uwe / Kardorff, Ernst von / Steinke Ines (eds.). (2000): *Qualitative Forschung: Ein Handbuch*. Reinbeck: Rowohlt.

Flick, Uwe (2002): *Qualitative Sozialforschung: Eine Einführung*. 6. überarb. und erweiterte Aufl. Hamburg: Rowohlt.

Giesecke, Michael (1991): *Der Buchdruck in der frühen Neuzeit: eine historische Fallstudie über die Durchsetzung neuer Informations- und Kommunikationstechnologien*. Frankfurt a. M.: Suhrkamp.

Keuneke, Susanne (2005): „Qualitatives Interview“, in: Mikos, Lothar / Wegener, Claudia (eds.): *Qualitative Medienforschung: Ein Handbuch*. Konstanz: UVK 254–267.

Kirschenbaum, Matthew (2008): *Mechanisms: New Media and the Forensic Imagination*. Cambridge: MIT University Press.

Manovich, Lev (2001): *Language of New Media*. Cambridge: MIT Press.

Polanyi, Michael (1985): *Implizites Wissen*. Frankfurt: Suhrkamp.

Przyborski, A. / Wohlrab-Sahr, M. (2009): *Qualitative Sozialforschung. Ein Arbeitsbuch*. München, Oldenbourg.

Stone, Sue (1982): „Humanities Scholars: Information Needs and Uses“, in: *Journal of Documentation* 38 (4): 292–313. http://www.emeraldinsight.com/journals.htm?articleid=1649976 .

Veit, Joachim (2010): „Es bleibt nichts, wie es war – Wechselwirkungen zwischen digitalen und ‚analogen‘ Editionen“, in: *editio* 24: 37–52 10.1515/9783110223163.0.37.

Warwick, Claire (2012): „Studying users in digital humanities“, in; Warwick, Claire / Terras, Melissa / Nyhan, Julianne (eds.), *Digital Humanities in Practice*. London: Facet Publishing 1–21.

Winter, Rainer (2010): „Handlungsmächtigkeit und technologische Lebensformen: Cultural Studies, digitale Medien und die Demokratisierung der Lebensverhältnisse“, in: Pietraß, Manuela / Funiok, Rüdiger (eds.) *Mensch und Medien. Philosophische und sozialwissenschaftliche Perspektiven*. Wiesbaden: VS 139–157.

3D-Metamodeling Christopher Polhem's *Laboratorium mechanicum* 1696

Snickars, Pelle

pelle.snickars@umu.se
Umea university, Schweden

In a letter from autumn 1696 to the Royal Swedish Bergs Collegium, the scientist and pre-industrial inventor, Christopher Polhem (1661-1751)—sometimes described as “the Father of Swedish Technology” (Lindroth 1951; Johnson 1963, Lindgren 2011)—argued that the Swedish king really ought to establish a *Laboratorium mechanicum*, all in order to foster future engineers. Importantly, this mechanical laboratory should have included educational *wood models* of contemporary equipment, machines and building structures, as well as water gates, hoistings and locks. Following Polhem, mechanics was simply the foundation of all knowledge: ”mechaniken är en grund och fundament til heela philosophien”. A few years later, a mechanical laboratory was indeed founded by Polhem, established near the Falu copper mine. Essentially his *Laboratorium mechanicum* became a pioneering facility (albeit small) for the pre-industrial training of Swedish engineers, as well as a laboratory for testing and exhibiting Polhem's own wooden models and designs. By the mid 1700, Polhem's *Laboratorium mechanicum* had transformed into the so called, *Royal Model Chamber*, a Swedish institution (funded by the king) for information and dissemination of technology and architecture set up in central Stockholm. It was admired, for example, by Johann Beckmann on his trip through Sweden in the mid 1760s. Later, during the 19th century, the pedagogical models belonging to the Royal Model Chamber were frequently used by engineering students at the KTH Royal Institute of Technology (in Stockholm). Apparently, this was especially the case with Polhem's so called *mechanical alphabet*. Initially, it consisted of 80 wooden models of basic machine elements like the lever, the wheel and the screw. Since a writer naturally had to know the alphabet in order to create words and sentences, Polhem argued that a contemporary *mechanicus* had to grasp his mechanical alphabet to be able to construct

and *understand* machines. Evidently, Polhem's models are interesting as physical traces of the *material foundations* of scientific knowledge (Ludwig, Weber & Zausig, 2014). Around 1930, however, part of the Royal Model Chamber and Polhem's mechanical alphabet collection was transferred to the Swedish National Museum of Science and Technology. Ever since it has served —and been frequently exhibited—as a kind of *meta-museological artifact*, since Polhem's designs proved to be pedagogical museological objects *avant la lettre*.

One of the objectives of the London Charter on computer-based visualisation of heritage promotes "intellectual and technical rigour in digital heritage visualisation" (London Charter 2009)—yet, in what way should one today digitise Polhem's *Laboratorium mechanicum*? What is the *exact* relation between "technical rigour" and virtual heritage in a software culture permeated by constant updates? Within the interdisciplinary Swedish research project, "Digital Models. Techno-historical collections, digital humanities & narratives of industrialisation" (funded by the Royal Swedish Academy of Letters, History and Antiquities) parts of Polhem's collection has been 3D scanned and 3D reconstructed by *different* software. The project set up is part of the trend were heritage institutions are today exploring how 3D technologies can broaden access to, and the understanding of their collections (Urban 2016; Ioannides 2014). Then again, is a 3D scan of a model (in our case) for example more *rigour* than a simulation?

In general, the research project "Digital Models" (that I am heading) explores the *potential* of digital technologies to reframe Swedish industrialisation and its stories about society, people and environments. The project uses three different cultural heritage perspectives to examine the *specificity of digitisation* and its potential to bridge research, institutional heritage and interest from the general public. Departing from the digitisation of three selected categories of material in the Swedish National Museum of Science and Technology collections, these mirror the three phases of industrialisation: (A.) parts of the business leader and industry historian, Carl Sahlin's extensive collection. (B.), all editions of the museum yearbook, *Daedalus* (1931-2014), and (C.) all of Polhem's preserved wooden models. These materials and phases correspond to three methodological approaches: traditional digitisation (A.), mass digitisation (B.) and critical digitisation (C.). Digitisation methods are hence correlated with different industrial-historical periods, resulting in three sets of digital tools, applications and/or game prototypes focused on various narratives of Swedish industrialisation.

In my presentation—done in English, but where questions can be posed in German since I am a fluent speaker—I want to present the ways in which we have worked with 3D modeling (parts of) Christopher Polhem's mechanical alphabet. Our 3D-metamodeling has been conceived as both a scholarly and as a museological practice. On the one hand we have tried to increase the historical understanding and knowledge about (and around) Polhem's models via visualisation, virtualisation and simulation, and on the other to experiment with novel ways to use the model's inherent pedagogical quality, and especially so within a museological context at the Swedish National Museum of Science and Technology. We have for example 3D scanned some of Polhem's models using a simple iPad iSense 3D scanner —and where we have also 3D printed some of our resulting imagistic models (with moving parts). Some of these digitisation activities have been performed within the actual museum space as a pedagogical activity, stressing the ways in which Polhem's old models still have a didactic quality to them. In addition, we have designed a few simple *virtual reality models* (of the models). Furthermore, in co-operation with Visualiseringscenter C (at Linköping University) we have also CT-scanned some of Polhem's models—i.e where images are taken from different angles to produce a cross-sectional and tomographic 3D image, a kind of virtual slice, allowing one to see inside the models without breaking them. Digital geometry processing has, in short, been used to generate a three-dimensional image of the inside of the models and their different parts. We have also co-operated with the professional animator Rolf Lindberg; on YouTube he has uploaded a number of videos of Polhem's models (Lindberg 2016). Lindberg, however, did not 3D scan Polhem's mechanical models—he *computer-animated* them in Cinema4D.

Hence, from a museological perspective, digitising Polhem's mechanical alphabet has produced a number of really different results. The London Charter on computer-based visualisation of heritage defines principles for the use of computer-based visualisation methods "in relation to intellectual integrity, reliability, documentation, sustainability and access" (London Charter 2009). Indeed, the charter recognises that the range of available computer-based visualisation methods is constantly increasing. Still, the linkage and

genealogy between copy and original sometimes becomes weak. For animator Lindberg, rather than 3D scanning Polhem's heritage items, it was way easier—and more pedagogical and visually enticing to *simulate* them—that is, building and constructing *brand new virtual objects*. The precious and highly esteemed original models collected at the museum—Polhem's mechanical alphabet—then becomes a model (rather than vice versa). Still, in the case of Polhem's models, the theme of (digital) reconstruction also has a profound historical dimension, since he sincerely believed (as a pre-industrial inventor) that *physical models* were always superior to drawings and abstract representations. The question is if he would have considered 3D reconstructions in a similar manner.

Bibliographie

Ioannides, Marinos et. al. (eds.) (2014): *Digital Heritage: Progress in Cultural Heritage*. Cham: Springer.

Johnson, William A. (1963): *Christopher Polhem, The Father of Swedish Technology*. Hartford: Trinity College Press.

Lindberg, Rolf (2016): https://www.youtube.com/channel/UC0UKj1XHuArjk-EZOrQRafQ .

Lindgren, Mikael H. (2011): *Christopher Polhems testamente*. Stockholm: Innovationshistoria förlag.

Ludwig, David / Weber, Cornelia / Zausig, Oliver (eds.) (2014): *Das Materielle Model*. Paderborn: Fink.

Lindroth, Sten (1951): *Christopher Polhem och Stora Kopparberget*. Uppsala: Almqvist & Wiksell.

London Charter (2009): http://www.londoncharter.org/fileadmin/templates/main/docs/london_charter_2_1_en.pdf .

Urban, Richard (2016): *Collections Cubed: Into the third dimension*. http://mw2016.museumsandtheweb.com/paper/collections-cubed-into-the-third- dimension/ .

Dokumentation, Werkzeugkasten, Pakete - Nachhaltigkeit von Daten und Funktionalität Digitaler Editionen

Czmiel, Alexander
czmiel@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Die Rolle digitaler Ressourcen in den Geisteswissenschaften und die zunehmende Bedeutung von Algorithmen werden noch immer unterschätzt. In den letzten Jahren bzw. Jahrzenten wurden zahlreiche Software-Werkzeuge, virtuelle Forschungsumgebungen und interaktive Publikationen, wie Datenbanken oder Digitale Editionen, für die geisteswissenschaftliche Forschung entwickelt. Diese werden innerhalb der Forschungscommunity mit zunehmender Tendenz akzeptiert und inzwischen breit eingesetzt. Jedoch stehen wir heute vor der Herausforderung diese verschiedenen Ressourcen, die zu einem großen Teil auf unterschiedlichen technischen Grundlagen basieren, weiter zu pflegen und verfügbar zu halten. Transparenz und Reproduzierbarkeit von Forschungsergebnissen, die mit einer digitalen Ressource oder Software erstellt wurden leiden darunter, dass diese Software oft nach wenigen Jahren nicht mehr gepflegt wird und damit nicht mehr lauffähig ist.

Der Vortrag beleuchtet die hier skizzierte Problematik am Beispiel Digitaler Editionen. Es werden drei Punkte vorgeschlagen, wie durch einen Bottom-Up-Ansatz die Nachhaltigkeit digitaler Ressourcen gefördert werden kann. Der Fokus liegt dabei auf den Erfahrungen, die im Laufe der letzten 10 Jahre mit der Entwicklung von XML-basierten Digitalen Editionen und dem Einsatz ausgewählter Software-Werkzeuge, wie der nativen XML-Datenbank eXistdb (http://exist-db.org), gewonnen wurden.

Bei digitalen Ressourcen handelt es sich um dynamische Objekte. Das bedeutet einerseits, dass die Inhalte jederzeit korrigiert, erweitert oder verändert werden können. Andererseits muss die technologische Basis fortlaufend aktuell, sicher und verfügbar gehalten werden. Diese beiden Prozesse sind Teilaufgaben eines

größeren Aufgabenbereichs, der unter dem Begriff „ *data curation*“ zusammengefasst werden kann.

Eine Digitale Edition ist mehr als nur ihre Forschungsdaten. Letztere, in vielen Fällen XML-Dokumente, werden oft erst durch eine adäquate Darstellung, durch Visualisierungen, wie Text-Bild-Verlinkungen, verschiedene Ansichten auf den Text, Netzwerke oder Timelines oder auch Verknüpfungen mit anderen externen Ressourcen „zum Leben erweckt“. Dieses Leben in Form programmierter Funktionalität ist ein genuiner Forschungsbereich der Digital Humanities. Allerdings führt die Funktionalitätsschicht (siehe Abbildung 1) einer Digitalen Edition auch dazu, dass *data curation* eine sehr komplexe Aufgabe werden kann, da hier die Flexibilität der Implementierung erheblich höher ist als auf der Ebene der Datenschicht.

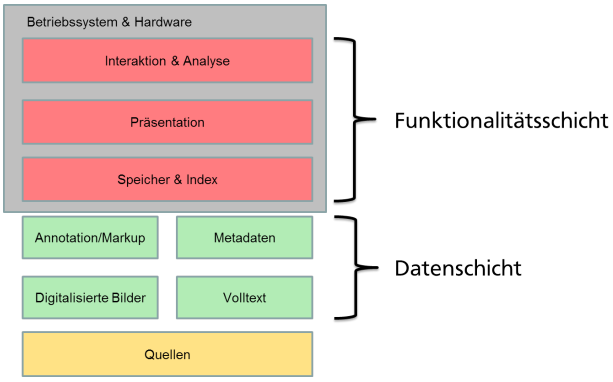


Abbildung 1: Schichtenmodell Digitale Edition

Digitale Editionen sind Software-Werkzeuge für die Analyse von Forschungsdaten. Damit sind sie ein Teil des Forschungsprozesses, der erhalten werden muss, um die Reproduzierbarkeit der Forschungsergebnisse zu gewährleisten.

Eine digitale geisteswissenschaftliche Ressource durchläuft üblicher Weise einen typischen Lebenszyklus. Dieser beginnt mit der Analyse der analogen Quellen, geht über die Datenmodellierung, die Auswahl bzw. Anpassung oder Neuentwicklung von Bearbeitungswerkzeugen sowie der digitalen Publikation der Forschungsergebnisse bis hin zu Fragen der Langzeitverfügbarkeit und Langzeitarchivierung. Bei jedem dieser Schritte sind verschiedenen Kompetenzen involviert. Das bedeutet, der Aufbau einer Digitalen Edition ist immer Teamwork. Dieses Team setzt sich in den meisten Fällen aus Personen zusammen, die einerseits das inhaltliche Fachwissen mitbringen und andererseits aus Personen mit einer Vielzahl unterschiedlicher technischer Kompetenzen:

Analyse der Quellen (Geisteswissenschaftler)
Anforderungsanalyse der digitalen Ressource, *Requirement Engineering* (alle Projekbeteiligten)
Entwurf des Daten- / Dokumentenmodells, Auswahl von Standards (Geisteswissenschaftler, Datenbankspezialisten, Markupspezialisten, Metadatenspezialisten)
Auswahl, Anpassung bzw. Entwicklung von Tools (Programmierer, Geisteswissenschaftler)
Aufsetzen und Betreuen der Server (Systemadministratoren)
Konzept, Design und Umsetzung der Web-Publikation (Webdesigner, Webentwickler, Geisteswissenschaftler)
Vorbereitung für Langzeitverfügbarkeit / -archivierung (Metadatenspezialisten, Dokumentationsspezialisten)
Betreuung und Wartung nach Projektende („ *data curators*“)

An jeder Stelle in diesem Lebenszyklus einer digitalen Ressource werden Entscheidungen getroffen, die Auswirkungen auf den nachfolgenden Schritt haben. So bilden die eigentliche Analyse der Inhalte, die z.B. in einer Digitalen Edition publiziert werden sollen, und die Anforderungsanalyse das Fundament, auf dem alles aufbaut, vom Daten- oder Dokumentenmodell, bis hin zur Publikation und der *data curation*.

Aus methodischer Sicht, mit besonderem Augenmerk auf das zugrundeliegende Text- bzw. Dokumentenmodell, wurden Digitale Editionen bereits ausführlich beschrieben (siehe Pierazzo 2015 und Sahle 2013). Eine Analyse aus technischer Sicht steht noch aus. Um die Entwicklung, Betreuung und Nachhaltigkeit Digitaler Editionen zu gewährleisten bedarf es eines technologischen Publikationskonzepts, das aus möglichst standardisierten Komponenten besteht.

Bisher existieren sehr erfolgreiche Standardisierungen auf dem Gebiet der Metadaten und der Textauszeichnungen, z.B. mit den Richtlinien der *Text Encoding Initiative* (TEI), aber wenig bis gar nichts bei der technischen Umsetzung und der Dokumentation. Dies würde helfen anschlussfähigere, stabilere und nachhaltigere digitale Ressourcen aufzubauen und damit auch die Arbeit eines Datenkurators, der sich um die Pflege dieser Ressourcen nach Projektende kümmert, deutlich vereinfachen.

Für eine aussichtsreiche Nachhaltigkeit kann die Lösung nicht allumfassend sein, sondern nur für klar definierte Anwendungsfälle gelten. In dem hier vorgestellten Fall ist dies eine XML-basierte Digitale Edition, die mit Technologien aus der X-Familie (XSLT, XQuery, XML-Schema, eXistdb) entwickelt wird. Das Grundprinzip ist jedoch auf andere Anwendungsszenarien übertragbar:

Eine ausführliche Dokumentation
Ein klar definierter Werkzeugkasten
Die Paketierung aller Projektressourcen

Dokumentation

Eine nachhaltige digitale Forschungsressource bzw. –software ist langfristig verfügbar, gut dokumentiert, lizenziert und versioniert, um die Reproduzierbarkeit der Forschungsprozesse zu garantieren. Die wichtigste Komponente ist eine ausführliche, formalisierte Dokumentation, die mindestens die folgenden Informationen enthalten sollte:

- Den Namen des Projekts und aller beteiligten Institutionen und Personen.
- Den Projektstatus: geplant, in Arbeit, veröffentlicht, beendet.
- Die eingesetzten Technologien und Standards inklusive Versionsangabe.
- Lizenzangaben zu Forschungsdaten, Quellcode, und anderen Komponenten, wie Schriftarten, Audio- oder Videodokumenten.
- Informationen darüber, wo der Quellcode und die Forschungsdaten zu finden sind.
- Informationen über die bereitgestellten APIs und andere Schnittstellen, um die Forschungsdaten in verschiedenen Formaten abzurufen (XML, HTML, PDF, JSON usw.) und in anderen Kontexten weiterzuverarbeiten.
- Details über die Forschungsmethode und den Hintergrund des Projekts. (Mehr dazu siehe Faniel 2015)
- Zitations- und Referenzierungsanweisungen für die persistente Adressierung aktueller und älterer Versionen der Forschungsdaten, Metadaten und Software.
- Eine standardisierte Historie der Projektentwicklung.

Selbstverständlich kann diese Liste nur ein erster Vorschlag sein. Sie enthält keinesfalls alle möglichen Informationen, die zu einer Digitalen Edition angegeben werden können. Die Dokumentation sollte maschinenlesbar (um z.B. als XML oder JSON weiter verarbeitet werden zu können) und über eine standardisierte Adresse bzw. einen klar definierten Zugriffspunkt (z.B. http://home.of.project/api/projectdescription) abrufbar sein. Dadurch wäre es möglich eine Digitale Edition bei einem zentralen Verzeichnis anzumelden, in dem alle Informationen und Updates über Digitale Editionen, die demselben Publikationsmodell folgen, gesammelt werden. Ein solches Verzeichnis existiert noch nicht.

Definierter Werkzeugkasten

Wie oben beschrieben kann Nachhaltigkeit nur in einem definierten Rahmen hergestellt werden, indem man klare Anwendungsfälle beschreibt. Selbst in diesen sind die Möglichkeiten der Umsetzung nahezu unbegrenzt. Daher ist es wichtig, genau zu definieren, welche Technologien, Standards und Software-Werkzeuge zum Einsatzkommen und welche Abhängigkeiten bestehen. Es ist ratsam die Zahl der eingesetzten Tools überschaubar zu halten und sich auf etablierte und gut dokumentierte Technologien zu konzentrieren.

Paketierung

Alle zusammengehörenden Komponenten einer Digitalen Ressource (Daten, Metadaten, Quellcode, Binärdatein, Dokumentation) müssen immer zusammen abrufbar sein. Diese Pakete tragen deutlich zu einer nachhaltigeren Entwicklung bei. Es ist immer klar, wo sich alle relevanten Informationen und Daten befinden. Zudem könnte ein Paket einer zentralen Kurationsstelle, z.B. einem Digital Huamnities Data Center, übergeben werden, die sich um die Betreuung der digitalen Ressourcen abgeschlossener Projekte kümmert. Diese Anlaufstelle existiert ebenfalls noch nicht.

Für den hier vorgestellten Anwendungsfall bietet das von eXistdb verwendete EXPath-Format (http://expath.org/) einen guten Ausgangspunkt für die Paketierung. Dieses Format beschreibt ein Packaging-System (http://expath.org/modules/pkg/), das es erlaubt XML-basierte Dokumente zusammen mit Abfrage- und Transformationsskripten sowie verschiedenen anderen Ressourcen auf eine standardisierte Art und Weise zu paketieren, das dieses Paket von allen Softwaresystemen, die diesem Standard folgen verstanden und ausgewertet werden kann. Damit kann dieses Packagesystem ähnlich fungieren, wie ein App-Store für Smartphones. Auch der Anpassungsaufwand für Software, die während der Projektlaufzeit eingesetzt wird, würde sich so verringern.

Andere Anwendungsszenarien lassen sich nicht auf Softwareebene paketieren. In diesen Fällen kann man auf Anwendungsvirtualisierung zurückgreifen,

wie sie zum Beispiel mittels Docker (https://www.docker.com/) möglich ist.

Um eine solche Standardisierung auf der technischen Ebene durchzuführen benötigt es eine aktive Digital-Humanities-Enwicklercommunity. Die Rolle des DH-Entwicklers ist im allgemeinen Diskurs noch deutlich unterrepräsentiert. Um wirklich erfolgreich Softwaretools für die geisteswissenschaftliche Forschung programmieren zu können, benötigt ein Entwickler mehr als nur ein grundlegendes Verständnis der typischen Problematiken in den einzelnen Fachdisziplinen. Umgekehrt erlauben Programmierkenntnisse ein wesentlich besseres Verständnis über die Funktionsweise der Software und damit bessere Einsatzmöglichkeiten sowie das Potential zu eigenen Verbesserungen. Um hier Fortschritte zu erzielen, müssen sich die DH-Entwickler besser organisieren. Die DH-Entwicklercommunity wäre der Kreis an Personen, die die Einführung und Anwendung von Standards, wie dem EXPath-System diskutieren und tragen. Mittelfristig wäre das Ziel für jeden Schritt im Lebenszyklus einer digitalen Ressource einen oder mehrere Vorschläge einer standardisierten Herangehensweise in Form eines *best-practice*-Leitfadens zu haben, der sich als *technical reader* zur Erstellung digitaler geisteswissenschaftlicher Ressourcen eignet und Vorschläge zu Schnittstellen, Standards, Lizenzen, Dokumentation, Zitationshinweise usw. anbietet.

Ein Ansatzpunkt dafür bietet die Arbeit des 2010 gegründeten Software Sustainability Institute (https://www.software.ac.uk/), Hong 2010 sowie Hettrick 2016, die verschiedene Ansätze zur Forschungs-Software-Nachhaltigkeit untersucht haben.

Bibliographie

Faniel, Ixchel (2015): *Data Management and Curation in 21st Century Archives*, 21. September 2015, http://hangingtogether.org/?p=5375 .

Hettrick, Simon (2016): *Research Software Sustainability. Report on a Knowledge Exchange Workshop*.

Hong, N. Chue et al. (2010): *Software Preservation Benefits Framework. Software Sustainability Institute Technical Report*.

Pierazzo, Elena (2015): *Digital Scholarly Editing, Theories, Models and Methods*. Ashgate.

Sahle, Patrick (2013): *Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*. 3 Bände. Norderstedt: Books on Demand.

Ein PoS-Tagger für „das“ Mittelhochdeutsche

Echelmeyer, Nora
nora.echelmeyer@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Reiter, Nils
nils.reiter@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Schulz, Sarah
sarah.schulz@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Einleitung

Ein grundlegender Schritt für eine Vielzahl von Aufgaben aus dem Bereich des *Natural Language Processing* (NLP) ist das *Part of Speech* (PoS)-Tagging. Ein PoS-Tagger annotiert im Kontext eines Satzes jedes Wort mit seiner Wortart aus einer Menge an festgelegten Wortarten (Tagset).

Ein Großteil der dazu vorhandenen Arbeiten konzentriert sich auf das Englische, auch für das Neuhochdeutsche sind vergleichbar viele Daten verfügbar. Historische Sprachstufen stellen hingegen eine Herausforderung für NLP-Aufgaben wie PoS-Tagging dar, da sie keine Standardsprache kennen, sondern nur als Vielfalt dialektaler Varietäten existieren, und ihre Verschriftlichung nicht nach einheitlichen Regeln erfolgt. Dies schlägt sich in einer hohen Varianz nieder, was die Annotation einer ausreichenden Menge an Referenzdaten erschwert.

Mit diesem Beitrag möchten wir einen PoS-Tagger für das Mittelhochdeutsche vorstellen, der auf einem thematisch breiten und diachronen Korpus trainiert wurde. Als Tagset verwenden wir ein Inventar aus 17 universellen Wortart-Kategorien (*Universal Dependency*-Tagset, Nivre et al. 2016). Mit den annotierten Daten entwickeln wir ein Modell für den TreeTagger (Schmid 1995), das frei zugänglich ist.

Dabei vergleichen wir drei verschiedene Möglichkeiten, den PoS-Tagger zu trainieren. Zunächst verwenden wir ein kleines, manuell

annotiertes Trainingsset, vergleichen dessen Ergebnisse dann mit einem kleinen, automatisch disambiguierten Trainingsset und schließlich mit den maximal verfügbaren Daten.

Mit dem Tagger möchten wir nicht nur eine „Marktlücke“ schließen (denn bisher gibt es keinen frei verwendbaren PoS-Tagger für das Mittelhochdeutsche), sondern auch eine größtmögliche Anwendbarkeit auf mittelhochdeutsche Texte verschiedener Gattungen, Jahrhunderte und regionaler Varietäten erreichen und weiteren Arbeiten mit mittelhochdeutschen Texten den Weg ebnen.

Forschungsstand

Tagset

Als PoS-Tagset hat sich für Neuhochdeutsch das Stuttgart-Tübingen-Tagset (STTS) etabliert (Schiller et al. 1999). Um auf die Besonderheiten historischer Sprachstufen besser eingehen zu können, entwickelten Dipper et al. (2013) ein hieran angelehntes Historisches Tagset (HiTS), das aus 12 Wortklassen besteht, die sich ihrerseits in 84 Wortarten gliedern.

Mit dem Ziel eines universellen Tagsets, welches konsistente Annotation vereinfacht und sprachübergreifendes Lernen für automatische Syntaxannotationen ermöglicht, wurde im Rahmen des *Universal Dependency*-Projekts (UD) ein Tagset aus 17 Tags erstellt. Dieses kann bei Bedarf um sprachspezifische Tags erweitert werden.

Tagging

Die besten verfügbaren PoS-Tagger erreichen auf englischsprachigen Zeitungstexten über 97% Accuracy (cf. Spoustová et al. 2009). Für deutschsprachige Zeitungstexte werden um die 95% erzielt, für Web-Texte 90–93% (Giesbrecht / Evert 2009). PoS-Tagging für das Mittelhochdeutsche ist weit weniger erforscht. Schulz / Kuhn (2016) beschreiben Ansätze zum PoS-Tagging eines spezifischen Textes. Barteld et al. (2015) trainieren einen PoS-Tagger für Mittelniederdeutsch. Dipper (2011) berichtet eine Accuracy von ca. 92% für zwei spezifische Modelle für die Dialekte Ober- und Mitteldeutsch, trainiert auf normalisierten Lemmata und mit dem STTS-Tagset. Alle genannten Modelle sind auf eine bestimmte Varietät des Mittelhochdeutschen beschränkt, zudem ist keines dieser Modelle (soweit uns bekannt) öffentlich verfügbar.

Korpora

Die wohl umfangreichsten Projekte zum Mittelhochdeutschen sind das Wörterbuchnetz [1], ein online zugänglicher Verbund aus Nachschlagewerken, das linguistisch motivierte Referenzkorpus Mittelhochdeutsch [2] (ReM, Dipper 2015) sowie die Mittelhochdeutsche Begriffsdatenbank [3] (s.u.).

Die annotierten Textkorpora (cf. Dipper 2015: 521–526) können z.T. über das Suchtool ANNIS (Zeldes et al. 2009) abgefragt werden, wobei Suchanfragen auf den Ebenen Wortform, Lemma und Morphologie möglich sind.

Mittelhochdeutsche Begriffsdatenbank (MHDBDB)

Durch eine Kooperation mit dem Projekt „Mittelhochdeutschen Begriffsdatenbank“ konnten wir für die Entwicklung unseres PoS-Taggers auf eine reiche Datensammlung zurückgreifen, bestehend aus 658 Texten mit insgesamt knapp 10 Millionen Tokens. Die Texte umfassen eine Zeitspanne von etwa vier Jahrhunderten (1100–1500), verschiedene dialektale Ausprägungen sowie nahezu alle Gattungen (von großepischen Genres wie Artusroman, Heldenepik und Antikenroman über Kleinepik hin zu Lyrik sowie diversen nicht-literarischen Texten wie Kochbüchern, Alchemistischen Schriften und Flugblättern).

Kürzel	Name	Beispiel
NOM	Nomen	acker, zît
NAM	Name	Uolrîch, Wiene, Rhîn
ADJ	Adjektiv	grôz, schoene
ADV	Adverb	schone, schnelleclîche
ART	Artikel	der, eine
DET	Determinante	ditze, mîn, ieman
POS	Possessiv-pronomen	mîn, dîn, unser
PRO	Pronomen	ich, ez, wir
PRP	Präposition	ûf, zuo, under
NEG	Negation	nie, âne, niht
NUM	Numeral	ein, zwô, zweinzegest
CNJ	Konjunktion	als, und, abr
GRA	Gradations-partikel	sêre, vil
IPA	Interrogativ-partikel	swer, swar, wie
VRB	Verb	liuhten, varn
VEX	Hilfsverb	haben, sîn, werden
VEM	Modalverb	müezen, suln
INJ	Interjektion	ahî, owê
CPA	Komparativ-partikel	als, wie
DIG	Zahl (Digit)	IX, XVII, III

Tabelle 1: Grammatische Kategorien der MHDBDB

Die Daten der MHDBDB enthalten – neben Tokenisierung und Lemmatisierung – bereits grammatische Auszeichnungen (Tabelle 1). Diese sind allerdings nicht disambiguiert, da sie den Kontext eines Wortes unberücksichtigt lassen (Typ-level-Annotationen, z.B. NOM|ADJ|ADV für *guot*). Darüber hinaus kodieren sie die morphologische Zusammensetzung von Wörtern (z.B. NOM|NEG für *unheil*), so dass es zu häufigen Mehrfachauszeichnungen kommt (z.B. *unvuoge* NOM|ADJ|ADV|NEG). Hinzu kommt, dass das Tagset nicht alle möglichen Verwendungsformen der Wörter abdeckt: So kann z.B. *daz* nicht nur Artikel oder subordinierende Konjunktion sein (Satz 1), sondern auch als Relativ- (2) oder Demonstrativpronomen (3) fungieren:

(1) Daz edel kint hât mir verjehen, daz ez in troume sî geschehen.
(2) Wie staete ist ein dünnez eis daz ougestheize sunnen hât?
(3) Daz sage ich iu vür ungelogen.

Trainings- und Testdaten

Obige Beobachtungen zeigen exemplarisch, dass die grammatischen Auszeichnungen der MHDBDB einer Überarbeitung bedürfen, um für die Entwicklung eines PoS-Taggers nutzbar zu sein. Dazu annotieren wir ein Teilkorpus, das für den mittelhochdeutschen Wortart-Tagger als Trainings- und Testdatei dient und für eine automatische Re-Annotation der restlichen MHDBDB-Daten herangezogen wird. Um Anschlussuntersuchungen sowie sprachübergreifende Betrachtungen zu ermöglichen, greifen wir für unsere Annotationen auf die universellen Kategorien aus dem UD-Tagset zurück (Tabelle 2).

Tag	Anmerkungen	Beispiele	
ADJ	adjective	vorangestellt, nachgestellt, Partizipien	der ritter **guot**; daz **elder** kint; **roemisch** lant; der **ander** man
ADP	adposition	Prä-, Post- und Zirkumposition	**mit** dem swerte; **gein** Nantes; **âne** ir schulde
ADV	adverb	auch adverbial gebrauchte Adjektive und relativischer Gebrauch	der ritter **lîdenlîche** leit; **so** sprach der künec; rehte liebe im **nie** geschach; hôret, **swie** er ze strîte quam
AUX	auxiliary verb	Hilfs- und Modalverben	ich **muoz** ir dienen; die **sint** enterbet; ir **habet** ez von mir gehôrt
CONJ	coordinating conjunction	nebenordnend	ritter **unde** diep; zwei teil **oder** mêr; **denne** ich welle jehen
DET	determiner	Artikel (bestimmt und unbestimmt); attribuierende Demonstrativ-, Possessiv- und Relativpronomen	**ein** maere; **der** ritter guot; **diz** bîspel; **dirre** âventiure; **ir** triuwe; **mîn** bruoder; **dehein** man; in **welhem** lande
INTJ	interjection		ouwê; ach!
NOUN	noun	auch substantivierte Adjektive, Verben, Numeralia	diu **vrouwe**; **duc** Orilus; **rîche** und **arme**; die **drî**; daz **singen**
NUM	numeral	nur Kardinalzahlen	die **drî** ritter
PART	particle	Negationspartikel; abgetrennter Verbzusatz; zu (mit Inf.); Vergleichspartikel; Abtönungspartikel	daz **en**weiz ich **niht**; lâzet allez trûren **abe**; daz ist swere **ze** halten; snêwîz **als** ein harm
PRON	pronoun	Personal-, Relativ-, Reflexivpronomen; substituierende Possessiv-, Indefinit-, Demonstrativ-, Interrogativpronomen	**er** lac tôt; Isenhârten, **der** den lîp verlôs; die kuenen heten **sich** berâten; der **sîne** sprach; da was **nieman**; **allez**, **daz** ich habe; **man** saget; **diz** was dô getân; **swaz** er gebôt
PROPN	proper noun	Eigennamen (auch mehrteilig)	Parzivâl; Orilus de Lalander; Nantes
SPUNCT	punctuation	Satzbeendende Zeichen	. : ! ?
PUNCT	punctuation	alle sonstigen Satzzeichen	, ; < > „ / ( ) usw.
SCONJ	subordinating conjunction	unterordnend	Er sagete **daz** Isenhart küneclîch bestatet wart; **sît** er an mir ist sus verzagt; **ob** mich gelücke wil bewarn
SYM	symbol		
VERB	verb	alle Vollverben	Er **lac** tôt; wir suln kurzwîl **phlegen**; er **hat** ein grôz her; ir reht was **vernomen**
X	other		

Tabelle 2: *Universal Dependency* (UD)-Tagset. Zum besseren Verständnis wurde das Tagset mit Beispielen und Anmerkungen versehen. Das Tag SYM wurde nicht benötigt; hingegen wurde das Tag SPUNCT hinzugefügt, um Satzbeendende Satzzeichen von anderen Satzzeichen zu unterscheiden.

Manuelle PoS-Annotationen

Das manuell annotierte Teilkorpus besteht aus 20.000 Tokens. Ein Teil der Daten (1.500 Tokens) wurde doppelt annotiert, um das Inter-Annotator-Agreement zu bestimmen (Cohen's kappa: 0.88; Cohen 1960). Um der Heterogenität der Sprache gerecht zu werden, enthält das Teilkorpus zufällig ausgewählte Abschnitte aus verschiedenen Textsorten des Gesamtkorpus.

Durch die Annotation aller Wörter im Kontext eines Satzes wurden Ambiguitäten aufgehoben. Zur Bestimmung der Wortart kann der Substitutionstest herangezogen werden, bei dem ein Wort durch ein Wort der gleichen Kategorie ersetzt wird. So wird *schoene* in *daz schoene wîp* durch ein anderes Adjektiv (z.B. *daz minnicliche wîp*), in *die schoene saz bî ime* hingegen durch ein Nomen ersetzt (z.B. *die vrouwe saz bî ime*).

Als Schwierigkeiten bei der Annotation stellten sich u.a. die Trennschärfe von DET und ADJ heraus (insb. für Wörter wie „viele"/ „alle") oder die Annotation von noch nicht lexikalisierten bzw. grammatikalisierten Formen (z.B. das mittelhochdeutsche *sît daz*, bei dem die Bestandteile ADP und PRON noch identifizierbar sind, wohingegen neuhochdeutsch „seitdem" eine SCONJ ist).

Ein weiterer Sonderfall des Mittelhochdeutschen besteht in der (weitgehend unsystematischen) Verwendung klitischer Formen, z.B. der Verschmelzung von Negationspartikel und Verb (*enmac*), der Kontraktion mehrerer Pronomen (*siz* = *sie* +*ez*), von Pronomen und Adposition (*zem* = *ze*+*im*) o.Ä. In solchen Fällen werden alle miteinander verschmolzenen Wörter annotiert, wobei ein + die Verschmelzung der Wörter anzeigt (*zem* ADP+PRON). Das UD-Tagset muss für das Mittelhochdeutsche also um „kombinierte Tags" (in unseren Daten finden sich 23 verschiedene Kombinationen) erweitert werden.

Automatische Disambiguierung des Gesamtkorpus

Das annotierte Teilkorpus dient neben seiner direkten Verwendung als Trainings- und Testkorpus (Modell 1) auch der automatischen Disambiguierung des Gesamtkorpus. Hierfür verwenden wir einen sequenziellen Tagger (Conditional Random Fields), der auf dem manuell annotierten Subkorpus trainiert wurde. Dieser lernt anhand der Annotationen und wortbasierten Eigenschaften, die ambigen Annotationen auf ihre disambiguierten Entsprechungen (UD-Tagset) abzubilden. Da sich in den Daten *auch* nicht-ambige Wörter befinden, lernt der Tagger an vielen Stellen 1-zu-1-Abbildungen, die als Anker fungieren können.

Die Disambiguierung des Gesamtkorpus erreicht eine Accuracy von 86,9%. Die auf diese Weise disambiguierten Daten kommen für die Modelle 2 und 3 als Trainingsdaten zum Einsatz.

Experiment und Evaluation

Um die Schwierigkeit der Aufgabe und die Tagging-Qualität einschätzen zu können, vergleichen wir drei verschiedene Modelle:

- Baseline: Anwendung des neuhochdeutschen TreeTagger-Modells auf den Testdaten.
- Modell 1: Der TreeTagger wird nur mit den manuell annotierten Daten trainiert, die Evaluation erfolgt als 5-fache Kreuzvalidierung (Cross-Validation), so dass in jedem Durchgang 16k Tokens als Trainingsdaten zur Verfügung stehen. Vorteil: Qualitativ hochwertige Trainingsdaten, Nachteil: Geringe Datenmenge.
- Modell 2: Der TreeTagger wird auf zufällig ausgewählten Sätzen trainiert, die zusammen etwa 16k automatisch disambiguierte Tokens umfassen. Die Trainingsmenge ist damit gleich groß wie für Modell 1 und erlaubt, die Auswirkungen der nicht-perfekten Disambiguierung abzuschätzen.
- Modell 3: Der TreeTagger wird mit allen automatisch disambiguierten Daten aus der MHDBDB trainiert (9,9M Tokens).

Modell	Precision	Recall	F-Score	Accuracy
Baseline	40,3	35,4	33,1	45,4
Modell 1 (kleines Trainingsset, manuell annotiert)	86,0	80,3	82,2	87,0
Modell 2 (kleines Trainingsset, autom. disambiguiert)	84,8	68,8	72,3	84,7
Modell 3 (großes Trainingsset, autom. disambiguiert)	91,2	79,6	82,9	90,9

Tabelle 3: Ergebnisse des PoS-Taggings mit verschiedenen Modellen. [4] Alle Modelle wurden auf den gleichen Daten evaluiert, für Modell 1 kam Cross-Validation zum Einsatz. Der Precision, Recall und F-Score ermöglichen eine tiefere Einsicht in die Performanz unter Berücksichtigung aller Wortartenklassen, während Accuracy die Gesamtperformanz sichtbar macht und als Vergleichswert zu State-of-the-Art-Ergebnissen dient.

Die Ergebnisse der unterschiedlichen Modelle sind in Tabelle 3 zusammengefasst. Zunächst zeigt sich erwartungsgemäß, dass die Baseline keine zufriedenstellenden Ergebnisse liefert. Modell 1 erreicht eine Accuracy von 87%, Modell 2 gut 2 Prozentpunkte weniger. Angesichts der Tatsache, dass die Trainingsdaten automatisch disambiguiert wurden, ist das nur ein geringer Verlust. Die Performanz steigt deutlich, wenn das große Datenset zum Training herangezogen wird (Modell 3). Gegenüber Modell 1 erreichen wir eine Verbesserung von ca. 3 Prozentpunkten Accuracy und damit insgesamt fast 91%. Eine Kombination der Modelle 1 und 3 erzielte keine Verbesserungen gegenüber Modell 3.

Eine Inspektion der von Modell 3 produzierten Annotationen ergibt, dass ein Großteil der Fehler (53%) auf die kombinierten Tags entfallen, die überwiegend als Pronomen oder Verben getaggt werden. Die nächsthäufigsten Fehlerklassen sind Numeralia und Partikeln. Die meisten Inhaltswörter (Nomen, Verben) werden korrekt erkannt (> 90%).

Fazit

Mit unserem Beitrag stellen wir einen nahezu universellen PoS-Tagger für das Mittelhochdeutsche vor, der auf Daten trainiert wurde, die dialektal, zeitlich sowie genremäßig variantenreich sind. Damit gehen wir davon aus, dass der Tagger auf ebensolchen Daten Ergebnisse erzielt, die ihn für darauf aufbauende Forschungen einsetzbar machen.

Daneben haben wir gezeigt, dass der Vorverarbeitungsschritt der Disambiguierung keineswegs perfekt funktionieren muss, um mit den Daten weiterzuarbeiten. Die Accuracy von ca. 87% für die Disambiguierung führt zwar bei gleicher Datenmenge zu einem Verlust an Tagging-Performanz, durch die größere, automatisch vorverarbeitete Datenmenge wird dieser aber mehr als aufgefangen.

Um die Nutzung unserer Forschungsergebnisse nachhaltig zu ermöglichen, stellen wir das Modell sowohl auf der TreeTagger-Webseite [5] als auch über eine Webanwendung [6] zur Verfügung. Des Weiteren ist das Modell als Ressource ins Clarin-D-Repositorium [7] aufgenommen, wodurch die Metadaten sowie die Links zum Modell permanent auffindbar bleiben.

Fußnoten

1. Online zugänglich unter http://woerterbuchnetz.de/ , letzter Zugriff 23.08.2016.
2. Online zugänglich unter http://referenzkorpus-mhd.uni-bonn.de , letzter Zugriff 23.08.2016.
3. Online zugänglich unter http://mhdbdb.sbg.ac.at/ (MHDBDB), letzter Zugriff 23.08.2016.
4. Für die Evaluation wurden die kombinierten Tags zu einer Klasse zusammengefasst.
5. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ .
6. www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/PoS_Tag_MHG.html .
7. Die Ressource kann im IMS-Repository gefunden werden: http://clarin04.ims.uni-stuttgart.de/fedora/objects/clarind-ims:92/datastreams/CMDI/content .

Bibliographie

Barteld, Fabian / Schröder, Ingrid / Zinsmeister, Heike (2015): „Unsupervised regularization of historical texts for POS

tagging“, in: *Proceedings of the 4th Workshop on Corpus-based Research in the Humanities (CRH)* 3–12 www.slm.uni-hamburg.de/germanistik/personen/zinsmeister/downloads/barteld-etai-2015.pdf [letzter Zugriff 23. August 2016].

Cohen, Jacob (1960): „A coefficient of agreement for nominal scales“, in: *Educational and Psychological Measurement* 20: 37–46.

Dipper, Stefanie (2015): „Annotierte Korpora für die Historische Syntaxforschung. Anwendungsbeispiele anhand des Referenzkorpus Mittelhochdeutsch“, in: *Zeitschrift für Germanistische Linguistik* 43: 516–563.

Dipper, Stefanie (2011): „Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison“, in: *Journal for Language Technology and Computational Linguistics* 26: 25–37 (= Proceedings of the TLT-Workshop on Annotation of Corpora for Research in the Humanities 2012) www.jlcl.org/2011_Heft2/2.pdf [letzter Zugriff 23. August 2016].

Dipper, Stefanie / Donhauser, Karin / Klein, Thomas / Linde, Sonja / Müller, Stefan / Wegera, Klaus-Peter (2013): „HiTS: ein Tagset für historische Sprachstufen des Deutschen“, in: *Journal for Language Technology and Computational Linguistics* 28: 85–137 www.jlcl.org/2013_Heft1/5Dipper.pdf [letzter Zugriff 23. August 2016].

Giesbrecht, Eugenie / Evert, Stefan (2009): „Is Part-of-speech tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus“, in: *Proceedings of the 5th Web as Corpus Workshop (WAC5)* www.stefan-evert.de/PUB/GiesbrechtEvert2009_Tagging.pdf [letzter Zugriff 23. August 2016].

Mittelhochdeutsche Begriffsdatenbank (MHDBDB). Universität Salzburg. Koordination: Margarete Springeth. Technische Leitung: Nikolaus Morocutti/Daniel Schlager. 1992–2016 http://mhdbdb.sbg.ac.at/ [letzter Zugriff 23. August 2016].

Nivre, Joakim / de Marneffe, Marie-Catherine / Ginter, Filip / Goldberg, Yoav / Hajič, Jan / Manning, Christopher D. / Mc Donald, Ryan / Petrov, Slav / Pyysalo, Sampo / Silveira, Natalia / Tsarfaty, Reut / Zeman, Daniel (2016): „Universal Dependencies v1: A Multilingual Treebank Collection“, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* 1659–1666 www.petrovi.de/data/lrec16.pdf [letzter Zugriff 23. August 2016].

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): „Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)“. Universität Stuttgart / Tübingen www.sfs.uni-tuebingen.de/resources/stts-1999.pdf [letzter Zugriff 23. August 2016].

Schmid, Helmut (1995): „Improvements in Part-of-Speech Tagging with an Application to German“, in: *Proceedings of the ACL SIGDAT-Workshop* 47–50 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf [letzter Zugriff 23. August 2016].

Schulz, Sarah / Kuhn, Jonas (2016): „Learning from Within? Comparing PoS Tagging Approaches for Historical Text“, in: *Proceedings of LREC 2010* 4316–4322 www.lrec-conf.org/proceedings/lrec2016/pdf/1237_Paper.pdf [letzter Zugriff 23. August 2016].

Entwicklung und Einrichtung einer digitalen Arbeitsumgebung für die *Jeremias Gotthelf-Edition*. Ein Erfahrungsbericht

Zihlmann, Patricia
patricia.zihlmann@germ.unibe.ch
Universität Bern, Forschungsstelle Jeremias Gotthelf, Schweiz

von Zimmermann, Christian
vonzimmermann@germ.unibe.ch
Universität Bern, Forschungsstelle Jeremias Gotthelf, Schweiz

Ausgangslage

Die *Historisch-kritische Gesamtausgabe der Werke und Briefe von Jeremias Gotthelf* (*HKG*), begründet von Prof. Dr. Barbara Mahlmann-Bauer und PD Dr. Christian von Zimmermann im Jahr 2003, ist auf 67 Text- und Kommentarbände angelegt und damit eines der grösseren Editionsvorhaben im deutschsprachigen Raum. [1]

Nach der Publikation der ersten Bände (2012) wurden die Arbeitsprozesse evaluiert, um die Edition adäquat auf digitale Arbeitsverfahren und Publikationsformen ausrichten zu können. Texte und Kommentare sollten nun *TEI*-konform

erfasst werden, zumal sich inzwischen die erweiterten *TEI-Guidelines* als internationaler editionsphilologischer Standard durchsetzen konnten. Die Umstellung sollte aber umfassender sämtliche editorischen Arbeiten von der Transkription und textgenetischen Analyse über die Dokument- und Datenverwaltung, die Registeranbindung oder Kontrollroutinen bis hin zur Publikation in Print und Web berücksichtigen. Ein wichtiges Anliegen war es, aus einem Datensatz mehrere Ausgabenformate erzeugen zu können (d.h.: unterschiedliche Printformate und digitale Präsentationsformen). Schliesslich sollten die 'endgültigen' Daten, die bisher allein in der historisch-kritischen Buchedition gedruckt vorlagen, in einem einheitlichen Datenformat in der *Forschungsstelle* zentral verfügbar, gesichert und anderweitig verwertbar sein. (Zuvor lagerten die durch Fahnenkorrekturen aktualisierten und in unterschiedlichen Satzprogrammen codierten Satzdaten bei mehreren beauftragten Satzbüros.)

Diese unterschiedlichen Aspekte machen das Umstellungsprojekt aussergewöhnlich komplex; einige der Etappenziele sind bereits erreicht worden. Unser Beitrag handelt von den Erfahrungen der vergangenen vier Jahre und gibt Aufschluss über Chancen und Schwierigkeiten des Umstiegs eines editorischen Grossprojekts auf ein exaktes, weitestgehend inhaltsorientiertes Markup und auf computerphilologische Arbeitsweisen. Von besonderem Interesse ist für uns die Vermittlung zwischen übergreifenden editorischen und digitalen Standards einerseits und individuellen Projektbedürfnissen andererseits. In unserem Vortrag möchten wir zudem den Einfluss institutioneller Rahmenbedingungen (*DaSCH/DDZ, SAGW, SNF, metagrid* [2]) auf die Durchführung computerphilologischer Reformen thematisieren.

Kooperation mit der BBAW und Pagina

Der Weg zu einer Arbeitslösung für die Edition erfolgt nicht durch eine eigene Programmentwicklung, deren Kosten auch bei einem nationalen Zusammenschluss mehrerer Projekte so nicht tragbar wären. Mittlerweile haben sich unterschiedliche modulare Lösungen etabliert (etwa auch *textgrid* etc.). Nach längerer Prüfung entschied sich die Projektgruppe dazu, die Arbeitsumgebung *Ediarum* der *BBAW* für das eigene Projekt weiterzuentwickeln, zumal *Ediarum* bereits für die *Schleiermacher-Edition* als digitale Arbeitsumgebung angewendet und auch für andere Projekte angepasst worden war. *Ediarum* basiert auf einer *eXist*-Datenbank, nutzt den *Oxygen Author* und umfasst Module für den Satzprozess sowie die digitale Datenpräsentation (Dumont/Fechner 2012; Arbeitsgruppe Telota o.J.).

Inzwischen sind erste Module unserer Arbeitsumgebung, die Schemata für Handschriften und Drucktexte nebst entsprechenden Anpassungen im *Oxygen Author* abgeschlossen. Für die Satzvorbereitung besteht eine Kooperation mit der Firma *Pagina*, welche das Satzmodul mit *Tustep* für die Druckvorstufe programmiert und in ersten Versionen für Drucktexte und Manuskripte bereits erfolgreich zur Verfügung gestellt hat (drucknaher Satzpreview für den Editionstext und Apparat).

Schema und Ausgabeformate

Die Entwicklung (insbesondere des Handschriftenschemas) war vor allem deshalb anspruchsvoll, da die Codierung nicht nur die Einrichtung der historisch-kritischen Ausgabe nach der bisherigen Gestalt sicherstellen musste, sondern auch davon abweichende gedruckte und digitale Präsentationsformen ermöglichen sollte. Dabei sollten etwa an die Stelle der Apparate der Drucktexte in der Webedition medienspezifische Annotationsformen treten können.

Die kritische Edition stellt einen Editionstext bereit und verzeichnet sämtliche textgenetischen Prozesse in der Handschrift, Emendationen bei Drucktexten und allfällige Varianten zwischen Drucktexten und Handschriften am Seitenende in unterschiedlichen Apparaten (textgenetischer Apparat, Emendationsapparat, Variantenapparat, Textstufenapparat; häufig Kombinationen).

Um die Apparate der kritischen Edition in der Printausgabe zu erzeugen, hätte ein Freitextelement wie <**note**> völlig ausgereicht. Doch hätte eine solche Lösung, welche die analytischen Stärken der Codierung für die editionsphilologische Arbeit nicht nutzt, den Aufwand für einen Umstieg auf *XML/TEI* nicht gerechtfertigt. Gerade die Möglichkeit zur präzisen Erfassung textgenetischer Prozesse einerseits und andererseits der Anspruch, einen Datensatz für unterschiedliche Ausgabeformate zu nutzen, legte den Umstieg nahe.

Digitale Präsentation und Korrespondenzedition

Die digitale Präsentation soll im Wesentlichen im Rahmen der Edition von Gotthelfs Korrespondenzen entwickelt werden (Zihlmann-Märki 2017). Dabei sind drei Ansichten für unterschiedliche Nutzungsszenarien vorgesehen. Neben der historisch-kritischen Ansicht, welche die *Tustep*-Routine einbindet, und einer diplomatischen Ansicht stellt eine Inhaltsansicht den finalen Text samt ausgezeichneten Entitäten und Stellenkommentaren bereit. Die Codierung wird die Vorschläge der *TEI Special Interest Group Correspondence* berücksichtigen, und die Verwendung standardisierter Daten ist Voraussetzung für eine Integration in externe Suchumgebungen wie *CorrespSearch*. [3] Ebenso können Informationen aus anderen digitalen Ressourcen dank dem Einsatz von Normdateien und dem BEACON-Format in der digitalen Umgebung angezeigt werden (Stadler 2012; Stadler 2014).

Vorzüge und Schwierigkeiten der Reform

Überblickt man die bisherige Reform, haben sich folgende Vorteile für die Arbeitsprozesse ergeben:

- Durch neue Arbeitsabläufe der Texterfassung (angepasste Scann- und OCR-Verfahren bei Drucktexten) konnte eine nicht unerhebliche Zeitersparnis erzielt werden.
- Unterschiedliche Previewansichten des Editionstextes erlauben die Hervorhebung spezifischer Besonderheiten, die für einzelne Korrekturschritte notwendig sind (etwa Hervorhebung des Zeilenfalles oder Hervorhebung von Stellen, die differenzierter codiert sind, als dies für die Druckedition im Apparat ausgegeben wird etc.).
- Die Dokumentation der Codierungsrichtlinien ermöglichte es, unterschiedliche Aspekte zu verknüpfen: 1) editorische Prinzipien konnten verbessert und verbindlicher gestaltet werden, 2) Probleme im Übergang von Transkription und Apparatgestaltung entfallen durch die Verbindung beider Prozesse, 3) die Verbindung von Codeerläuterung, Transkriptionsbeispiel und Präsentationsbeispiel hat der Satzfirma die Programmierung der Satzroutinen erleichtert.
- Die Satzfirma *Pagina* investiert vor allem in die Konzeption von Satzroutinen und kommt dann nur noch für den Feinsatz der Buchausgabe zum Einsatz. Dank Roundtripping können die Satzinformationen (Seiten- und Zeilenzahlen) in die originalen *XML*-Daten zurückgespielt werden; so verfügt die *Forschungsstelle* über aktuelle Daten, die alle Informationen zur Druckausgabe enthalten und leicht auch für andere Präsentationsformen oder – in weiteren Reformetappen – für digitale Querverweise und Kommentarverankerungen genutzt werden können.

Zugleich erwies sich der Reformprozess als überaus anspruchsvoll sowie als zeit- und kostenintensiv.

Als Illusion erweist sich die Vorstellung, es sei möglich, eine Codierung unabhängig von späteren Ausgabeformaten zu entwickeln. Editionen, welche die Daten tatsächlich für mehrere Formate bereit halten wollen und nicht nur auf eine spezifische Ausgabe zielen, stehen hier vor bedeutenden Problemen, da sie die Dateninterpretation durch Webapplikationen ebenso berücksichtigen müssen wie die Eigenheiten von Satzroutinen oder den Wunsch nach einer mediumsspezifischen Apparatgestaltung. Dies gilt umso mehr für Projekte, welche die Umstellung im laufenden Arbeitsprozess bei bereits etablierten Editionsrichtlinien durchführen.

Innerhalb der modularen Arbeitsumgebung konnte aufgrund der projektspezifischen Bedürfnisse letztlich kein einziges Modul unverändert übernommen werden. Da die Module jeweils spezifischen Projektinteressen der an der Modulentwicklung beteiligten Projektpartner folgen (müssen), sind sie – auch nach Erfahrung von *Telota* – in keinem Fall ohne Anpassung nutzbar. Andere Module (Druckvorstufe) konnten dagegen problemlos ausgetauscht werden, und dies wäre wohl eine Grundanforderung überhaupt an modulare Editionsumgebungen.

Die graphische Oberfläche in *Ediarum* können wir für unsere Edition nicht nutzen, weil in ihr komplexe Textphänomene und sich überlagernde Korrekturen nicht übersichtlich darstellbar sind. Allerdings hat sich auch gezeigt, dass

die Arbeit in der Code-Ansicht für die meisten Mitarbeitenden unproblematisch ist. Für einzelne Arbeitsschritte wie die Lemmatisierung wäre die Arbeit in der graphischen Oberfläche möglich.

Allein von der Kostenseite aus betrachtet, lohnt sich bei begonnenen Buchausgaben die Umstellung auf eine differenzierte Codierungspraxis nur bei einem bedeutenden Projektvolumen. Erst dann stehen die Entwicklungs- und Anpassungskosten einer digitalen Arbeitsweise in einem Verhältnis zur Einsparung von Satzkosten. Das gilt auch dann, wenn auf eine Buchedition gänzlich verzichtet wird.

Standardisierung und heterogene Editionslandschaft

Freilich wäre es zwecks Ressourcenschonung wünschenswert, dass andere Projekte von Schemata, von Satzroutinen, von der Arbeitsumgebung und der digitalen Präsentation profitieren könnten, die in unserem oder einem anderen Projekt entwickelt wurden. Tatsächlich sehen wir ein gewisses Potential im Erfahrungsaustausch. Dass eigentliche Übernahmen von Projektstrukturen hingegen schwierig sind, liegt weniger an den Codierungsstandards als an heterogenen Editionstypen und -prinzipien. Verbindliche Prinzipien der Textwiedergabe und Apparatierung könnten die Digitalisierungsprozesse wie auch die langfristige Sicherung vereinfachen (*ein* Prozedere, *ein* Umwandlungsschema für alle Daten) und möglicherweise kostensenkend wirken. Die Diversität von Editionen entspringt aber nicht zufälligen Entwicklungen, sondern ist tief in heterogenen Forschungstraditionen verankert, und die Entscheidung für einen Editionstyp geht in der Regel mit einer intensiven Auseinandersetzung mit dem Editionsgegenstand einher. So rechtfertigt die Berner Parzival-Edition ihr Projekt nicht zuletzt durch Annahmen über den mittelalterlichen Textbegriff (Stolz 2002), und die Berner Gotthelf-Edition hebt auf den politisch-publizistischen wie diskursiven Charakter der Texte ab, der nur durch eine umfassende Kommentierung adäquat dargestellt werden kann (von Zimmermann 2014).

Auch die *TEI* trägt der Diversität prinzipiell Rechnung, stellt sie doch einen Pool möglicher Codes zur Verfügung, aus denen die Einzelprojekte ihre eigenen Schemata erarbeiten müssen. Ein über das Bekenntnis zu *TEI/XML* (oder einer anderen Auszeichnungssprache) hinausgehender, projektübergreifender Standard, der die textphilologische Kernarbeit (Transkription und Apparatierung) betrifft, kann deshalb vermutlich eher nicht entwickelt werden.

Fußnoten

1. Allgemeine Projektinformationen: http://www.gotthelf.unibe.ch
2. http://www.metagrid.ch
3. http://correspsearch.net/index.xql

Bibliographie

Dumont, Stefan / Fechner, Martin (2012): *Digitale Arbeitsumgebung für das Editionsvorhaben „Schleiermacher in Berlin 1808–1834“*. http://digiversity.net/2012/digitale-arbeitsumgebung-fur-das-editionsvorhaben-schleiermacher-in-berlin-1808-1834 [letzter Zugriff 25. August 2016].

[**Arbeitsgruppe Telota**]: *Ediarum – Digitale Arbeitsumgebung für Editionsvorhaben*. http://www.bbaw.de/telota/software/ediarum [letzter Zugriff 25. August 2016].

Schweizerische Akademie der Geistes- und Sozialwissenschaften (2015): *Final report for the pilot project „Data and Service Center for the Humanities“ (DaSCH)*. Swiss Academies Reports 10 (1).

Stadler, Peter (2012): „Normdateien in der Edition“, in: *editio* 26: 174–183.

Stadler, Peter (2014): „Interoperabilität von digitalen Briefeditionen“, in: Delf von Wolzogen, Hanna / Falk, Rainer (Hg.): *Fontanes Briefe editiert*. Internationale wissenschaftliche Tagung des Theodor-Fontane-Archivs Potsdam, 18. bis 20. September 2013 (= Fontaneana 12). Würzburg: Königshausen & Neumann 278–287.

Stolz, Michael (2002): „Wolframs ‚Parzival‘ als unfester Text. Möglichkeiten einer überlieferungsgeschichtlichen Edition im Spannungsfeld traditioneller Textkritik und elektronischer Darstellung“, in: Haubrichs, Wolfgang / Lutz, Eckart C. / Ridder, Klaus (Hg.): *Wolfram von Eschenbach – Bilanzen und Perspektiven*.Eichstätter Colloquium 2000 (= Wolfram-Studien 17). Berlin: Schmidt 294–321.

Zihlmann-Märki, Patricia (2017): „Kommentierung in gedruckten und digitalen Briefausgaben“, in: Lukas, Wolfgang / Richter, Elke (Hg.): *Kommentieren und Erläutern im*

digitalen Kontext (= Beihefte zu editio) [erscheint 2017].

von Zimmermann, Christian (2014): „Geschichte, Ziele und Perspektiven der Historisch-kritischen Gesamtausgabe der Werke und Briefe von Jeremias Gotthelf (HKG)“, in: Marianne Derron / Christian von Zimmermann (Hg.): *Jeremias Gotthelf*. Neue Studien. Hildesheim / Zürich / New York: Olms 13–37.

Hermann Burgers *Lokalbericht*: Hybrid-Edition mit digitalem Schwerpunkt

Daengeli, Peter
p.daengeli@uni-koeln.de
Cologne Center for eHumanities

Zumsteg, Simon
szumsteg@vtxmail.ch
Schweizerisches Literaturarchiv

Einleitung

Dreht sich ein Roman vorab um das Roman-Schreiben, den Zustand der zeitgenössischen Literatur und die Kritik an ihr, dann ist es vielleicht gar nicht so abwegig, dass sein cleverster Witz in seiner eigenen Nicht-Veröffentlichung steckt. Der schweizerische Schriftsteller Hermann Burger griff auf den letzten Seiten seines Romans *Lokalbericht* (1970) zu genau dieser selbstreflexiven und auto-dekonstruktiven Volte, als er den Mentor des jungen Protagonisten und angehenden Schriftstellers urteilen lässt, so könne man heute nicht mehr schreiben, das Manuskript solle liegen bleiben, “ein Jahr, zwei Jahre, zehn Jahre lang” – und Burger sich offenbar selbst an den Ratschlag hält und den Roman tatsächlich zeit seines Lebens nicht veröffentlicht.

Ausschlaggebend für die Nicht-Veröffentlichung war freilich nicht diese Pointe, sondern ein biographischer Umstand (vgl. dazu den Kommentar in Zumsteg 2016a: 257-304 bzw. http://www.lokalbericht.ch/kommentar). Aus heutiger Warte erscheint die Veröffentlichung jedoch zweifellos geboten. Der *Lokalbericht* nimmt in Burgers Lebenswerk und Werkleben eine Scharnierfunktion ein, indem sich Burger in diesem ‘Rohdiamanten’ erstmals ungestüm an jene unverwechselbare Poetik herantastet, die ab seinem Roman *Schilten: Schulbericht zuhanden der Inspektorenkonferenz* von 1976 zu seinem Markenzeichen wird. Die lang währende Beschäftigung Burgers mit dem Romantext und seinen Vorstufen – schreibend, vorlesend, auszugsweise publizierend – legt darüber hinaus Zeugnis ab für die Bedeutung, die er dem Text beimaß. Burgers Bonmot “Literatur ist, wenn man trotzdem druckt”, muss also zur Rechtfertigung dieser postumen Veröffentlichung gar nicht erst bemüht werden.

Der Lokalbericht als Hybrid-Edition

Die erstmalige Herausgabe des Romans aus dem Nachlass im Schweizerischen Literaturarchiv, Bern (SLA), erforderte eine andere editorische Handhabe als die vor zwei Jahren erschienene Leseausgabe von Burgers Werken in acht Bänden (Zumsteg 2014), die sich auf den Wiederabdruck seiner bereits zuvor publizierten Texte beschränkte. Eine Hybrid-Edition, bestehend aus einer umfassenden digitalen Edition und einem schlichten Lesebändchen als deren Spin-off (Zumsteg 2016a, 2016b), wird dem Archivfund besser gerecht, zumal sich die Entstehung des Romans dadurch auszeichnet, dass einzelne Textbausteine eine lange Vorgeschichte haben. Die Änderungen und Unterschiede werden dabei viel weniger auf den einzelnen Typoskript-Seiten manifest als zwischen den Textstufen, sie werden erst im Vergleich der betreffenden Dokumente erkennbar. Die Möglichkeiten einer digitalen Edition, die nicht durch das Format der Buchseite und die lineare Folge bestimmt und begrenzt sind, boten sich für diese Situation besonders an. Das digitale Format gab überdies Gelegenheit, die vielfältigen allographen Materialien aus dem Nachlass und aus anderer Provenienz miteinzubeziehen, die für die Kontextualisierung und das Verständnis der Autographen sowie für Burgers mosaikartige Arbeitsweise erhellend sind. Das Projekt bot also eine gute Ausgangslage, den vielgerühmten Mehrwert einer digitalen Edition gegenüber den hergebrachten Publikationsformen zu realisieren. Die digitale Edition (beta) ist seit dem 22. Oktober 2016 unter http://www.lokalbericht.ch verfügbar.

Grundlage: hochqualitative Digitalisate, dokument-orientierte TEI-Encodings

Neben den 179 beschriebenen Typoskriptseiten des Romans wurden von den knapp 900 weiteren Textträgern, die im Rahmen der digitalen Edition präsentiert werden, hochauflösende Digitalisate im TIFF-Format und/oder mit OCR hinterlegte PDF-Dateien angefertigt. Aufgrund der fotografisch hervorragenden Qualität und der ohnehin guten Leserlichkeit der (zumeist nur punktuell korrigierten) Typoskripte und handschriftlichen Dokumente nehmen diese Digitalisate den Rang der primären digitalen Repräsentation ein.

Die Digitalisate aller Dokumente, die Bestandteil des eigentlichen dossier génétique sind, wurden durch dokument-orientierte, auf der Basis von OCR-Daten erstellte, TEI-Encodings (sourceDoc) ergänzt. Diese Encodings bilden die Grundlage der diplomatischen Transkription, d.h. der Umschrift aller Texte inklusive mikrogenetischer Varianz. Die minutiöse Aufzeichnung dieser Phänomene erwies sich als sehr zeitaufwändig. Retrospektiv wäre zu erwägen, die Tiefe des Encodings im Sinne des Vermeidens eines “Over-taggings” (Bernhart/ Hahn 2014, Hanrahan 2015) zu reduzieren, zumal die derart codierten Phänomene aus der Perspektive der naheliegenden Forschungsfragen nur einen begrenzten Mehrwert gegenüber dem durchsuchbaren Volltext und den Digitalisaten schaffen.

Durch Extraktion der Text-Nodes inklusive der regularisierten Varianten und basierend auf Milestone-Elementen (Absätze, Seitenumbrüche) ließen sich ab dem dokument-orientierten TEI-Encoding sowohl der für die Druckausgabe verwendete Lesetext als auch text-orientierte TEI-Encodings gewinnen, welche die Grundlage für die Lesefassung der digitalen Edition bilden (Umschrift aller Texte ohne Berücksichtigung mikrogenetischer Varianz).

Die TEI-Encodings, die unter den Bedingungen der CC BY 2.0-Lizenz nachnutzbar sind, enthalten selbstredend auch die Metadaten des jeweiligen Typoskript-Konvoluts oder Briefes.

Präsentation und Funktionalität

Als digitale Edition mit ausgeprägter textgenetischer Komponente lehnt sich der Aufbau der *Lokalbericht*-Edition an vergleichbare aktuelle Projekte wie die textgenetische Edition Wofgang Koeppens *Jugend* (Krüger, Mengaldo, Schumacher 2016) und die historisch-kritische Edition von Goethes *Faust* an (Bohnenkamp, Henke, Jannidis 2016). Die angebotenen Interaktionsmechanismen lassen sich mit der Edition Ch. G. Heynes' Vorlesungen über die Archäologie vergleichen (Graepler o. J.). Die Benutzerin oder der Benutzer der Edition hat in jeder Dokumentansicht die Möglichkeit, zwischen den vier Präsentationsformen Digitalisat, Transkription, Lesefassung und Metadaten hin und her zu wechseln. Dabei gibt es drei Ansichtsmodi, die für unterschiedliche Lese- bzw. Benutzungspraktiken stehen:

In der **Grundansicht** wird eine Oberfläche eines Textträgers dargestellt. Schaltflächen erlauben das Navigieren innerhalb des Texts, offerieren aber auch Verknüpfungen zu textgenetisch verwandten Seiten anderer Konvolute. Die verlinkten Ziele werden standardmäßig an Stelle des aktuellen Dokumentes geladen. Die Grundansicht eignet sich daher besonders für eine (zumeist) lineare Lektüre. Sie lässt sich im Gegensatz zu den beiden Doppelansichten auch auf kleinen Anzeigegeräten verwenden.

Die **Parallelansicht** dient dazu, die gleiche Textstelle in zwei Ansichten zu vergleichen. Sie erlaubt z.B. die Gegenüberstellung von Digitalisat und Transkription. Vor- und Zurückblättern wirkt sich dabei jeweils auf beide Ansichten aus. Auch in der Parallelansicht lassen sich Seiten anderer Konvolute laden, die einen textgenetischen Bezug zur aktuell dargestellten Seite haben.

Die flexibelste Benutzungsmöglichkeit bietet schließlich die **Synopsis**. Sie ähnelt beim ersten Aufruf der Parallelansicht, lässt im Unterschied zu ihr aber den Vergleich zwischen textgenetisch verwandten Textträgern Seite an Seite zu. In dieser Ansicht wird typischerweise zweimal die gleiche Präsentationsform gewählt, etwa Faksimile gegen Faksimile oder diplomatische Transkription gegen diplomatische Transkription.

Digitale Edition *Lokalbericht*: Synoptische Darstellung

Alle drei Ansichtsmodi binden unter der Textansicht auch eine visuelle Navigation in der Form eines skizzenbasierten Graphs ein, der eine Vogelperspektive auf das gesamte Materialkorpus bietet. Indem in Benutzerinteraktion die textgenetischen Bezüge als Verbindungslinien (Links) zwischen den als Kreissymbolen (Nodes) dargestellten Einzelblättern ein- und ausgeblendet werden können, lässt sich die Textentwicklung auf der Makroebene anschaulich verfolgen. Die einzelnen Verbindungslinien dienen dabei zugleich als Links, über die sich die beiden ausgewählten Blätter wiederum in der synoptischen Ansicht laden lassen (Dängeli, Theisen, Wieland, Zumsteg 2016).

Editoriale Expertise zur Textgenese wird zusätzlich durch einen erläuternden Überblickskommentar und Stellenkommentare zum Romantext sowie durch Verweise auf editionsinterne und -externe Ressourcen befördert.

Durch die Auszeichnung von Personen (421), Orten (378) und Werken (191), die im Roman bzw. im Korpus vorkommen oder einen wichtigen Bezug dazu haben, ist das Korpus überdies auch semantisch erschlossen. Diese wo immer möglich auf Normdatensätze (VIAF, GND, GeoNames) referierenden Einheiten sind über Register zugänglich, sie lassen sich aber auch in der Transkription und im Lesetext hervorheben.

Die Volltextsuche mit flexibel kombinierbaren Filtern (z.B. nach Textkategorie) bietet dem Benutzer einen weiteren Einstiegspunkt in die vielfältigen und spannenden Materialien der digitalen Edition.

Aspekte digitaler Nachhaltigkeit

Weil von Anfang an die Perspektive bestand, die digitale Edition nach der Entwicklung am Cologne Center for eHumanities an das Schweizerische Literaturarchiv bzw. die Schweizerische Nationalbibliothek zu übergeben –, galt ab Projektbeginn die Prämisse, eine bewusst einfache und leichtgewichtige technische Lösung anzustreben, die nur einen geringen Anteil an serverseitiger Progammierung erfordert und die sich mittel- und langfristig leicht warten lässt. Diese Kriterien legten es nahe, aus den TEI-Encodings statisches HTML zu erzeugen, das nur punktuell durch dynamisch erzeugten Code ergänzt werden muss (z.B. freie Volltextsuche). Dass eine solche datenbanklose Lösung durchaus zeitgemäß sein kann, zeigen die zahlreichen Generatoren für statische Webseiten, die in jüngerer Zeit entwickelt wurden und die mitunter große Akzeptanz fanden. [1]

Für die digitale Lokalbericht-Edition fiel die Wahl mit Apache Cocoon [2] auf eine vom Ansatz her in gewissem Grad vergleichbare, jedoch besser zu den vorliegenden XML-Daten passende und insbesondere auch sehr ausgereifte Anwendung. Dabei profitierten wir von der soliden Grundlage des im Produktiveinsatz bewährten Werkzeugs *Kiln* (vormals *xMod*) von Monteiro Viera und Norrish (2012), das Cocoon mit SOLR und Sesame bündelt und dank guter Dokumentation schnell einsatzbereit ist. [3] Ergänzt um eigene Pipeline-Definitionen und XSL-Transformationen ließ sich auf der Grundlage von Kiln eine monolithische Webanwendung erstellen, die abgesehen vom Bildserver alle Funktionalitäten umschließt und die mit einem einzigen Befehl auf einem Standardwebserver lauffähig ist. Im Bedarfsfall lässt sich die Seite auch als komplett statisches HTML abspeichern, wodurch ein wesentlicher

Schritt zur langfristigen Konservierung und Verfügbarmachung erfüllt sein sollte.

Um die Hürde der Applikations-Präservierung auch hinsichtlich der permanenten Referenzierung tief zu halten, verweisen die Permalinks lediglich auf – irgendwie geartete – digitale Repräsentationen real existierender Dokumente. Bestimmte Ansichten oder Funktionalitätszustände sind damit explizit nicht permanent referenzierbar, was die Verantwortung der übernehmenden Institution reduziert und ihr mehr Flexibilität für künftige konzeptuelle oder technische Veränderungen zubilligt. Im Bedarfsfall sind die Benutzer gehalten, bestimmte Ansichten selbständig zu persistieren, beispielsweise durch Übergabe der URL an die Wayback Machine. [4]

Ein weiterer Mosaikstein zur Sicherung der Nachhaltigkeit betrifft die (derzeit laufende) Aufnahme der Ressource durch das Data Center for the Humanities (DCH) [5] der Universität zu Köln, in deren Rahmen neben der Klärung technischer und rechtlicher Fragen auch die periodische Prüfung der Ressource nach festgelegten Kriterien geregelt wird. Möge dies gewährleisten, dass der zu Lebzeiten unpubliziert gebliebene Lokalbericht nach seiner Erstveröffentlichung nicht abermals ins archivalische Dunkel versinkt.

Fußnoten

1. Die ihrerseits mit DocPad erstellte Liste unter https://staticsitegenerators.net führt per November 2016 445 derartige Tools auf. Zur Beliebtheit vgl. auch https://www.staticgen.com. Die Vorteile dieses Ansatzes liegen auf der Inputseite vorab in den einfachen Quellformaten, die zumeist zur Verwendung kommen (z.B. Markdown, Textile, YAML), auf der Outputseite in der durch die direkte HTML-Auslieferung bedingten hervorragenden Performanz (Biilmann Christensen 2015, Kraetke/Imsiek 2016, Rinaldi 2015).
2. Vgl. https://cocoon.apache.org/2.1/.
3. Vgl. zu Kiln auch Turska 2014.
4. Die Umsetzung einer technischen Adressierbarkeit granulärer Dateneinheiten war nicht Bestandteil des Projekts, sie könnte auf der Bestehenden Grundlage aber nachgerüstet werden.
5. Vgl. http://dch.phil-fak.uni-koeln.de.

Bibliographie

Bernhart, Toni / Hahn, Carolin (2014): „Datenmodellierung in digitalen Briefeditionen und ihre interpretatorische Leistung. Ontologien, Textgenetik und Visualisierungsstrategien. Workshop im Jacob-und-Wilhelm-Grimm-Zentrum der Humboldt-Universität zu Berlin, 15./16. Mai 2014“, in: *editio* 28: 225-229.

Biilmann Christensen, Mathias (2015): „Why Static Website Generators Are The Next Big Thing“, in: *Smashing Magazine* 2. November 2015 https://www.smashingmagazine.com/2015/11/modern-static-website-generators-next-big-thing/ [letzter Zugriff 24. August 2016].

Bohnenkamp, Anne / Henke, Silke / Jannidis, Fotis (2016): *Historisch-kritische Faustedition*. Unter Mitarbeit von Gerrit Brüning, Katrin Henzel, Christoph Leijser, Gregor Middell, Dietmar Pravida, Thorsten Vitt und Moritz Wissenbach. Beta-Version 2. http://beta.faustedition.net [letzter Zugriff 28. Oktober 2016].

Daengeli, Peter / Theisen, Christian / Wieland, Magnus / Zumsteg, Simon (2016): „Visualizing the Gradual Production of a Text“, in: *DH2016: Conference Abstracts* 767–769.

Graepler, Daniel (o. J.): *Christian Gottlob Heyne – Vorlesungen über die Archäologie*. http://heyne-digital.de [letzter Zugriff 20. November 2016].

Hanrahan, Elise (2015): „‚Over-tagging‘ with XML in Digital Scholarly Editions“, in: *DHd 2015: Von Daten zu Erkenntnissen* 193–196.

Kraetke, Martin / Imsieke, Gerrit (2016): „XSLT as a powerful static website generator. Hogrefe's Clinical Handbook of Psychotropic Drugs“, in: *Proceedings of XML In, Web Out: International Symposium on sub rosa XML. Balisage Series on Markup Technologies* 18 http://www.balisage.net/Proceedings/vol18/html/Kraetke02/BalisageVol18-Kraetke02.html [letzter Zugriff 24. August 2016].

Krüger, Katharina / Mengaldo, Elisabetta / Schumacher, Eckhard (2016): *Wolfgang Koeppen*. Jugend. http://www.koeppen-jugend.de/ [letzter Zugriff 24. August 2016].

Monteiro Viera, Jose Miguel / Norrish, Jamie (2012): *Kiln*. https://github.com/kcl-ddh/kiln [letzter Zugriff 24. August 2016].

Rinaldi, Brian (2015): *Static Site Generators*. Modern Tools for Static Website Development. Sebastopol: O’Reilly.

Turska, Magdalena (2014): „What prevents people from firing their own Kiln?“, in: *Nesting Instinct. Build in progress* http://blogs.it.ox.ac.uk/mturska/2014/07/30/what-prevents-people-from-

firing-their-own-kiln [letzter Zugriff 24. August 2016].

Zumsteg, Simon (2014): *Hermann Burger: Werke in acht Bänden*. Zürich: Nagel und Kimche.

Zumsteg, Simon (2016a): *Hermann Burger – Lokalbericht: Roman. Herausgegeben aus dem Nachlass*. Zürich: Voldemeer.

Zumsteg, Simon / Dängeli, Peter / Wieland, Magnus / Wirtz, Irmgard (2016b): *Hermann Burger – Lokalbericht: Digitale Edition*. http://www.lokalbericht.ch [Beta-Version vom 22. Oktober 2016].

Kontextbasierte Zitationsanalyse soziologischer Klassiker im Verlauf von 100 Jahren

Messerschmidt, Reinhard
reinhard.messerschmidt@uni-koeln.de
Universität zu Köln, Deutschland

Mathiak, Brigitte
mathiak@gmail.com
Universität zu Köln, Deutschland

Bibliometrische Zitationsanalyse ist in den Naturwissenschaften allgemein üblich geworden, in den Geistes- und Sozialwissenschaften jedoch mit Problemen hinsichtlich der Datenbasis und unterschiedlicher Zitationsweisen konfrontiert. So betonen Sula und Miller (2013), dass verschiedene Referenzkontexte nicht ignoriert werden dürfen, da intellektuelle Dispute zum geisteswissenschaftlichen Kern gehören. Für drei Klassiker der Soziologie haben wir daher die Zitationen in ihrem Zitationskontext, sowie im zeitlichen Verlauf analysiert und verschiedene gängige Hypothesen zu Trends in diesem Bereich statistisch überprüft.

Datenbasis und Methode

Das zentrale Textkorpus besteht aus digitalisierten Tagungsbänden („Verhandlungen“) der Deutschen Gesellschaft für Soziologie (DGS) von 1910 bis 2010 und umfasst 6869 Dokumente, sowohl direkt konvertiert aus dem Ausgangsmaterial, als auch OCR-behandelte Scans. Beide Dokumenttypen wurden zunächst nach diktionär- und n-gram basierter Vorbereitung (Saad und Mathiak 2013) in reinen Text konvertiert. Wie in den Geisteswissenschaften üblich, sind zwar viele wichtige Akteure im Korpus präsent, aber deren Hauptwerke üblicherweise Monografien, für die zusätzliche Quellkorpora konstruiert wurden. Ausgewählt wurden mit Karl Marx, Max Weber und Theodor W. Adorno drei Klassiker, deren gesammelte bzw. ausgewählte Schriften digital in hoher Qualität vorliegen und für die Soziologie selbst eine entscheidende Rolle spielen. Weber ist dabei mit Abstand der am häufigsten zitierte und gilt als Ahnherr der deutschen Soziologie. Marx wurde (gemeinsam mit Friedrich Engels) stark und insbesondere auch kontrovers diskutiert. Dem Werk Adornos kommt in der Soziologie und Sozialphilosophie der 1960er Jahre eine herausragende Stellung zu. Aufgrund des selektiven Charakters des DGS-Korpus bezüglich jeweiliger Tagungsthemen wurde zur Ergänzung ein Korpus aus seit 1949 digital verfügbaren Fachzeitschriften [1] erstellt und annotiert.

Der Fokus des Projekts lag auf der Analyse von Text-ReUse sowie Sentiments in Zitationen und Paraphrasen. Vorhandene Ontologien wie CiTO [2] erwiesen sich aufgrund vieler für unsere Zwecke irrelevanter Kategorien als zu komplex und zeitaufwändig. Aus Effizienzgründen und auch um Erkenntnisse in Bezug auf Sentimentpolarität (Boland et al. 2013) nutzen zu können, haben wir uns primär auf drei Ausprägungen letzterer konzentriert: positiv, negativ und neutral. Weitere im Rahmen von Sentimentanalysen übliche Differenzierungen hinsichtlich z.B. gradueller Abstufung und Subjektivität (Pang und Lee 2008) wurden bewusst ausgeblendet. Nach ersten Annotationsversuchen wurde allerdings klar, dass zusätzliche Kategorien für Ambivalenz und Negationsstrukturen notwendig sind. Insgesamt 3382 Codes [3] wurden dabei interpretativ und kontextbezogen von soziologischen Experten in MaxQDA [4] auf Basis des zuvor beschriebenen Codeschemas annotiert. Im Verlauf dieses Prozesses wurden spezifische Charakteristika des Korpus deutlich, die bei der Analyse zu berücksichtigen waren:

Erstens zeigten sich im Datensatz strukturelle Brüche hinsichtlich der Dokumentanzahl je Tagung. Während diese von 1910 bis 1979 bereits zwischen 9 und 94 variiert, existieren ab 1980 abgesehen von einer Ausnahme zwischen 240 und 675 Dokumente. Dadurch sind absolute Zahlen von Autorennamen nicht vergleichbar (Abb. 1).

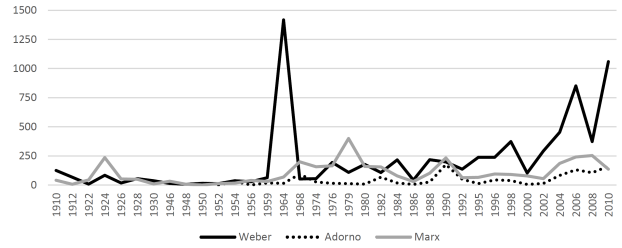


Abbildung 1: absolute Ergebnisse pro Jahrgang (Maximum bei Weber 1964 bedingt durch auf ihn bezogenes Thema der Tagung, Einbruch im Jahr 2000 dagegen durch außergewöhnlich niedrige Zahl von Dokumenten, siehe Abb. 4)

Zweitens ist, wie Sula und Miller (2013) bereits betonten, die Abgrenzung einer Zitation insbesondere in älteren Dokumente nicht immer klar. Die zunächst simple Keyword-Suche nach Autorennamen führt zu systematischer Überschätzung der Referenzen aufgrund 1) Erwähnung von Autoren in anderem Kontext z.B. von Zusammenfassungen wie „klassische Autoren (insbesondere Simmel, aber auch Marx und Weber)“ oder 2) Biografischen Darstellungen sowie 3) Literaturverzeichnissen. Zusätzlich zeigt sich insbesondere bei Weber das Problem der Autorendisambiguierung, denn nicht jeder Weber ist Max: angefangen von seinem Bruder Alfred über seine Frau Marianne bis zu insgesamt 30 weiteren – teilweise in denselben Dokumenten. Zusätzlich müssen unterschiedliche Zitationsstile sowie die parallele Zitation mehrerer Werke berücksichtigt werden, da andernfalls eine Unterschätzung vorliegt. Spezifische Abkürzungen wie z.B. siebzig Mal „MWG“ für Webers gesammelte Schriften oder 102-fach „MEW“ für ausgewählte Werke von Marx/ Engels und vergleichsweise weit von der Erwähnung des Autorennamens positionierte Referenzen würden nur allzu leicht übersehen, wenn nicht auch explizit nach diesen gesucht wird.

Dies wirft einerseits die Frage nach der angemessenen Definition einer Referenz im Kontext dieses Projekts auf sowie andererseits danach, welche entsprechende Darstellung im zeitlichen Verlauf adäquat ist. Hinsichtlich ersterer fiel die Entscheidung zugunsten maximierter Offenheit und Inklusion, was insbesondere bei Weber, welcher oft nur mittels Erwähnung des Namens ohne genauen Werkbezug referenziert wird, viele Fälle kanonischer Zitation einschließt und sogar Fälle indirekter Zitation, in welchen andere Autoren referenziert wurden z.B. „in Webers Terminologie (vgl. Habermas 1982)“. Der Grund dafür ist, dass andernfalls unter ausschließlicher Berücksichtigung nur formal korrekter Referenzen deren Fallzahl massiv abnimmt (Abb. 2).

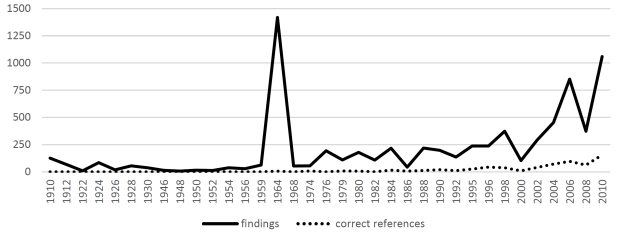


Abbildung 2: absolute Ergebnisse pro Jahrgang vs. Formal korrekte Referenzen für Weber

Im Fall Max Webers stehen nur 640 korrekte Referenzen insgesamt 7381 Suchergebnissen gegenüber, was einem Verlust von 91,3% entspricht, welcher insbesondere darauf hinausläuft, dass alle Referenzen von den 1970er Jahren auf Grund formaler Defizite entsprechend heutiger Standards verloren gehen. Daran wird die Bedeutung sogar der lockersten Erwähnungen sichtbar. Die zweite zuvor aufgeworfene Frage nach der adäquaten Repräsentation erwies sich als kompliziert und wird im folgenden Abschnitt diskutiert.

Bibliometrische und wissenschaftsgeschichtliche Ergebnisse

Im langfristigen Trend liefe die Interpretation von Abbildung 2, abgesehen vom themenbedingten Ausreißer des Jahres 1964, auf ein zunächst zurückgehendes und dann eine in den 1960er und 70er Jahren anfangs langsame Renaissance hinaus, welche sich im neuen Jahrhundert intensiviert. Dieser Effekt einer sogenannten Weber-Renaissance (Glassman 1983, Hinz 1966) ist auch aus der Fachliteratur bekannt. Werden die Ergebnisse aber in Relation zu den Dokumenten pro Tagung berechnet, wird es schwieriger, von einer solchen Renaissance zu sprechen, wenngleich der Ausreißer von 1964 bleibt (Abb. 3).

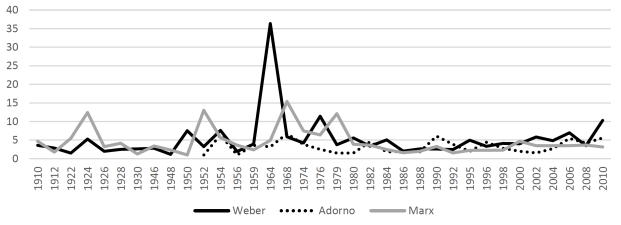


Abbildung 3: durchschnittliche Referenzen pro Dokument (Verhältnis Ergebnisse/ Anzahl der ergebnisbeinhaltenden Dokumente)

Beim Blick auf die Dokumentanzahl pro Jahrgang (Abb. 4) zeigt sich die eingangs erwähnte Heterogenität sowie ein generell zunehmender Trend, was den Unterschied der relativen zur absoluten Darstellung erklärt.

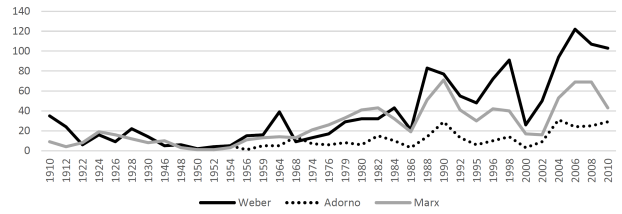


Abbildung 4: absolute Dokumentanzahl pro Jahrgang

Wenn weiterhin diese absolute Dokumentanzahl pro Jahrgang ins Verhältnis zur Gesamtzahl jährlicher Dokumente gesetzt wird (Abb. 5), zeigt sich vielmehr ein abnehmender Trend für alle Autoren.

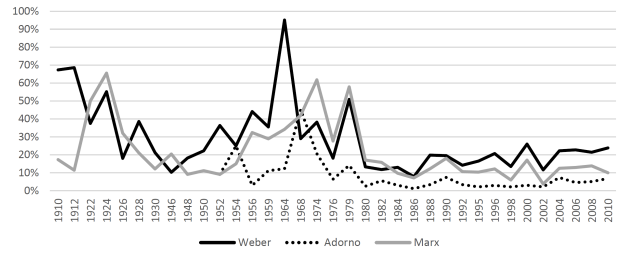
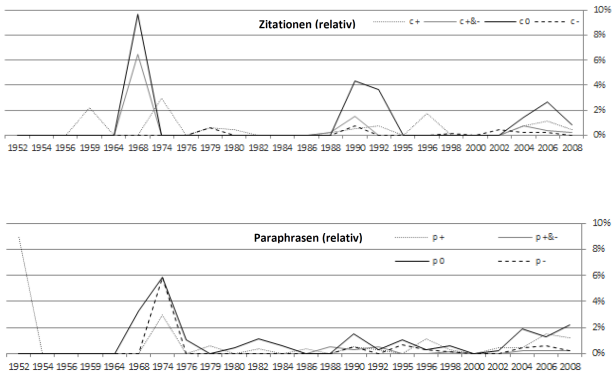


Abbildung 5: Prozentualer Anteil an Dokumenten pro Jahrgang (Verhältnis Anzahl der ergebnisbeinhaltenden Dokumente/ Gesamtzahl der Dokumente jeweils eines Jahrgangs)

Alle gezeigten Grafiken sind Konstruktionen unter Hervorhebung verschiedener Aspekte, wenngleich die letzten beiden hinsichtlich der Intensität zu verschiedenen Jahrgängen informativer erscheinen als die vorangegangen. Trotz der immer noch vagen Verbindung jeweiliger Zahlen von Referenzen angesichts der zuvor dargestellten Definitionsprobleme, lässt sich jedoch kein guter Grund für die Präferenz einer der in den letzten beiden Abbildungen dargestellten Berechnung finden. Nichtsdestotrotz zeigen beide Optionen [5] keinesfalls eine Weber-Renaissance.

Demgegenüber zeigt sich bei der für Adorno durchgeführten vertieften kontext- und sentimentbezogenen Analyse ein klar abnehmender Trend. Angesichts der im Vergleich zu Weber viel geringeren Anzahl an Referenzen welche erst ab 1952 auftreten konnten wir die Sentimentpolarität (positiv, ambivalent, neutral, negativ) von Zitationen (Abb. 6) und Paraphrasen (Abb. 7) detailliert annotieren.



Abbildungen 6, 7: Sentimentpolarität für Zitationen und Paraphrasen

(relative Häufigkeiten in Bezug auf Dokumentanzahl pro Jahrgang)

Zunächst zeigen sich in beiden Abbildungen lokale Maxima um 1968, welche angesichts Adornos enormer Rezeption im Kontext der 68er-Bewegung kaum überrascht – im Gegensatz zur leichten Verzögerung bei Paraphrasen, welche möglicherweise durch die zunehmende Bekanntheit seiner Werke bedingt ist. Möglicherweise könnte ein solches Schema generell im Hinblick auf die Entstehung zukünftiger Klassiker auftauchen, was genauer zu untersuchen wäre. Die zweite überraschende Beobachtung besteht in der trotz damals hochgradiger Polarisierung der Disziplin (z.B. im „Positivismusstreit“) starken Häufigkeit neutraler Referenzen. Diese ist jedoch vielfach durch eine spezifische Argumentationsstruktur bedingt, in der nach vielen neutral-deskriptiven Aussagen letztendlich nur wenige polarisierte verwendet werden.

Abschließend können wir berichten, dass die Schwierigkeiten, Geisteswissenschaften durch die „positivistische“ bibliometrische Tradition adäquat abzubilden auch konzeptuell real sind und nicht nur der schwierigen Datenlage angelastet werden können. Ein mehr an derartiger Analyse, die sich jedoch nur auf die kleinen Einheiten der Auseinandersetzung konzentriert, verfehlt das Gesamtbild. Trotzdem kann sie als Hilfsmittel eingesetzt werden, um neue Wege der wissenschaftsgeschichtlichen Annäherung zu eröffnen, darunter insbesondere die Auseinandersetzung im distant reading (Moretti 2013).

Fußnoten

1. Soziale Welt, Kölner Zeitschrift für Soziologie und Sozialpsychologie, Deutsche Zeitschrift für Philosophie
2. http://purl.org/spar/cito/
3. http://cceh.uni-koeln.de/share/annotation_soc_classics.zip
4. http://www.maxqda.de/
5. Die in der Bibliometrie übliche auf Textlänge basierende Berechnung erwies sich angesichts diesbezüglicher Heterogenität des DGS-Korpus als nicht anwendbar.

Bibliographie

Adorno, Theodor W. (2004): „Theodor W. Adorno, Gesammelte Schriften“, in: *Digitale Bibliothek* 97.

Boland, Katarina/ Wira-Alam, Andias/ Messerschmidt, Reinhard (2013): „Creating an annotated corpus for sentiment analysis of German product reviews“, in: *GESIS-Technical Reports* 2013/05 http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2013/TechnicalReport_2013-05.pdf [letzter Zugriff 19. August 2016].

Glassman, Ronald (1983): „The Weber renaissance“, in: *Current Perspectives in Social Theory* 4: 239–271.

Hinz, Horst (1966): „Max-Weber-Renaissance?“, in: *Vierteljahreshefte zur Wirtschaftsforschung* 4: 454–479.

Marx, Karl / Engels, Friedrich (2004): „Marx, Engels, ausgewählte Werke“, in: *Digitale Bibliothek* 11.

Moretti, Franco (2013): *Distant Reading*. London: Verso.

Pang, Bo / Lee, Lillian (2008): „4.1.2 Subjectivity Detection and Opinion Identification“, in: *Opinion Mining and Sentiment Analysis* . Now Publishers Inc.

Sula, Chris Alen / Miller, Matt (2013): „Citation studies in the humanities“, in: *DH2013: Conference Abstracts* http://dh2013.unl.edu/abstracts/ab-353.html [letzter Zugriff 19. August 2016].

Saad, Farag / Mathiak, Brigitte (2013): „Revised mutual information approach for german text sentiment classification“, in: *WWW '13 Companion. Proceedings of the 22nd international conference on World Wide Web* 579–586. http://dl.acm.org/citation.cfm?id=2487788.2487997 [letzter Zugriff 19. August 2016].

Weber, Max (2004): „Max Weber, gesammelte Werke“, in: *Digitale Bibliothek* 58.

Langzeitinterpretierbarkeit auf Basis des CIDOC-CRM in inter- und transdisziplinären Forschungsprojekten am Germanischen Nationalmuseum (GNM), Nürnberg

Große, Peggy
p.grosse@gnm.de
Germanisches Nationalmuseum, Deutschland

Wagner, Sarah
s.wagner@gnm.de
Germanisches Nationalmuseum, Deutschland; für MUSICES[1]

Im musealen Bereich spielt die Frage, wie man langfristig interpretierbare Daten erzeugt und bereitstellt, eine immer größere Rolle, insbesondere wenn, wie am Germanischen Nationalmuseum (GNM), drittmittelgeförderte inter- und transdisziplinäre Forschungsprojekte große Datenmengen zu den Objektbeständen erheben. Welche Lösungsansätze für den nachhaltigen Umgang mit Forschungsdaten das GNM verfolgt, soll anhand zweier Forschungsprojekte dargestellt werden.

1. Anforderungen und Ziele des transdisziplinären Forschungsprojektes zu Friedensrepräsentationen in der Vormoderne

Das von der Leibniz-Gemeinschaft seit Juli 2015 geförderte internationale Kooperationsprojekt „Repräsentationen des Friedens im vormodernen Europa“ erforscht Friedensbilder im Zeitraum vom 16. bis 18. Jahrhundert. Friedensvereinbarungen mussten über den reinen Vertragstext hinaus erklärt, begründet und vermittelt werden. Das übernahmen Friedensrepräsentationen, die ein multimediales Phänomen der Frühen Neuzeit waren. Folglich nimmt das Forschungsprojekt visuelle Darstellungen, sprachliche Bilder sowie musikalische Ausprägungsformen in den Blick.

Dieser breite Ansatz erfordert die Kooperation unterschiedlicher geisteswissenschaftlicher Fachrichtungen mit ihren jeweiligen Analysekompetenzen und Perspektiven sowie Institutionen mit geeigneten Beständen. [2]

Um abstrakte Konzepte wie Frieden, Gerechtigkeit oder Wohlstand darzustellen, verwendeten Künstler, Dichter oder Komponisten einen Kanon von Motiven, die europaweit genutzt und verstanden wurden. Dieses „Vokabular" des Friedens soll beispielhaft erschlossen und über die Gattungs- und Genregrenzen hinweg analysiert werden. Zudem wurden gemeinsame Fragestellungen zu transmedialen Rezeptionsvorgängen, Veränderungen der Motivik im Zusammenhang mit unterschiedlichen Friedensschlüssen, zu Funktion und Wahrnehmung von visuellen, sprachlichen und musikalischen Konzepten entwickelt. Am Anfang steht daher die transdisziplinäre Erfassung und Nutzung der heterogenen Bestände.

Ein entsprechendes Dokumentationssystem muss demzufolge für alle beteiligten WissenschaftlerInnen unabhängig von der Art und Darstellungsform der Quellen gewährleisten, dass sie schnell und effizient die relevanten Informationen eingeben und abrufen können. Die erfassten Informationen beziehen sich auf objektbezogene Daten, aber auch auf deren Inhalte und Form, wie Ikonografie, Textgattung oder Instrumentierung. Außerdem soll der inhaltliche Zusammenhang zwischen Objekten und Friedensereignissen dokumentiert werden. Daher muss die Datenbank in der Lage sein, auch Zusammenhänge strukturiert abbilden zu können. Die Ergebnisse sollen in einem virtuellen Themenportal am Ende des Projektes veröffentlicht werden. Die Einbindung von digitalen Bild-, Text- und Musikquellen ist daher wünschenswert, ebenso die Möglichkeit mit einem Thesaurus arbeiten und bereits vorhandene Normdaten einbinden zu können.

2. Anforderungen und Ziele des interdisziplinären Forschungsprojektes MUSICES

Das Projekt „MUSICES" (MUSIkinstrumenten-Computertomographie-Examinierungs-Standard) hat es sich zur Aufgabe gemacht einen Standard zu entwickeln, der die Bedingungen für eine wissenschaftliche und praxisnahe Abbildung von Musikinstrumenten durch 3D-Computertomographie beschreibt. Das zerstörungsfreie, bildgebende Verfahren der Computertomographie ist ein wichtiges Instrument geworden, um Informationen über den Aufbau und die Konstruktion von Musikinstrumenten zu gewinnen und so Aussagen über Herstellungsweise, Erhaltungszustand und klangliche Eigenschaften zu liefern. In Kooperation von WissenschaftlerInnen und RestauratorInnen des Germanischen Nationalmuseums und des Fraunhofer Instituts EZRT (Entwicklungszentrum Röntgentechnik) in Fürth werden gemeinsam die technischen Parameter, effiziente und objektschonende Praxisabläufe sowie die Möglichkeiten und Grenzen dieser Technik intensiv erarbeitet.

Die Entwicklung des Standards besteht aus verschiedenen Aspekten: Zunächst bedarf es eines Schemas, das den kompletten Ablauf der Untersuchung des Instruments dokumentiert, von der Auswahl eines Objekts und die Fragestellung an dieses bis über den Transport, die eigentliche Messung und deren Parameter sowie die daraus erzeugten 3D-Röntgenbilder. Im Laufe des Projekts werden über 100 verschiedene Instrumente erforscht, die in ihrer Auswahl eine möglichst große Vielfalt an Eigenschaften abbilden sollen, um die Anwendbarkeit des Standards auch auf andere Objekte übertragen zu können. Unterschiedliche Materialien und die geometrischen Formen der Musikinstrumente spielen bei den einzustellenden Parametern der 3D-CT eine entscheidende Rolle, um die gewünschten Resultate zu erzielen. Für die Objekte werden deshalb ihren Eigenschaften entsprechend Kategorien definiert. Auf diese Weise können Richtwerte entwickelt werden, beispielsweise für die Strahlungsdosis, die vom Material und dessen Stärke abhängig sind. Die Relation zwischen Objektkategorie und Messeinstellungen in Abhängigkeit von der Forschungsfrage muss durch das Dokumentationsschema abgebildet werden. Letzteres muss zudem aufgrund der stetigen Optimierung des Untersuchungsprozesses während des Projektverlaufs flexibel gestaltet sein.

Als Teil des Standards soll das Dokumentationsschema in bestehende Standards integriert werden und als Metadatenmodell für künftige Projekte dienen, die sich mit der 3D-CT von Objekten beschäftigen. Alle gewonnenen Daten sollen zum Projektende in das Objektdokumentationssystem des Germanischen Nationalmuseums integriert, darüber hinaus aber auch an internationale Portale geliefert und öffentlich zugänglich gemacht werden.

3. CIDOC CRM und WissKI als Werkzeuge der Dokumentation und Langzeitinterpretierbarkeit

Die semantische Erschließung, die eine nachhaltige Interpretierbarkeit von heterogenen

Forschungsdaten zunächst innerhalb einer Institution gewährleistet, erfolgt auf Grundlage einer Ontologie, die es ermöglicht Wissen formal zu definieren, zu kategorisieren, zu beschreiben und auszutauschen. Forschungsprojekte am GNM verwenden das ISO-zertifizierte Conceptual Reference Model (CIDOC CRM, ISO 21127, Doerr / Lampe / Krause 2011). [3] Da das CRM nicht maschinell lesbar ist, wurde dies im sog. „Erlangen-CRM“ [4] auf Basis von OWL [5] nachgeholt (Görz 2011).

Damit die Projektdaten in einem gemeinsamen Kontext unter Verwendung einer gemeinsamen „Sprache“ dokumentiert werden können, werden Anwendungs- bzw. Domänenontologie, basierend auf dem CIDOC CRM, für jedes Projekt entwickelt. Der Austausch von Daten und deren Langzeitinterpretierbarkeit wird durch die gemeinsame Basis des CIDOC CRM gewährleistet, während alle Spezifika der jeweiligen Projekte möglichst fachspeziell durch die Domänenontologie abgedeckt sind (Hohmann / Fichtner 2015, 117-118). Dies geschieht unter dem Vorbehalt, dass innerhalb einer Institution die Klassen und Eigenschaften gleich gehandhabt werden.

Um das angesprochene kollaborative und transdisziplinäre Arbeiten zu ermöglichen, benötigt man eine virtuelle Forschungsumgebung. Ausgewählt wurde WissKI [6], dessen Fokus auf dem interaktiven und vernetzten Arbeiten basierend auf semantischer Tiefenerschließung mit Hilfe des Erlangen-CRM liegt. Die Erfassung kann text- und formularbasiert erfolgen. Die Oberflächen des Systems können den jeweiligen Bedürfnissen der Projekte angepasst werden, wobei die Form der Wissensrepräsentation und die Wiederverwendung der Daten gattungs- und disziplinübergreifend ermöglicht wird. Darüber hinaus können digitale Bild-, Text- und Audiodateien angezeigt und verwaltet werden. Zudem unterstützt WissKI die Erstellung lokaler Vokabulare und die Nutzung bestehender Normdaten.

3.1 Anwendungsbeispiel Projekt „Friedensrepräsentationen“

Das zentrale Anliegen des Projektes zur Analyse der Friedensrepräsentationen ist ein transdisziplinärer und vergleichender Forschungsansatz basierend auf einer kooperativen Erschließung und Nutzung heterogener Quellenbestände. Angaben zu den Objekten und ihren Inhalten müssen ebenso wie historische Daten zu Friedensereignissen erfasst werden. Diese unterschiedlichen Informationen sollen semantisch vernetzt sein, um eine langfristige und nachhaltige Interpretierbarkeit sicher zu stellen. Eine Herausforderung ist es, spezifische Daten und Anforderungen unterschiedlicher Fachdisziplinen zu vereinheitlichen und Schnittpunkte zu bilden. Das CIDOC CRM erlaubt durch die Definition geeigneter übergeordneter Abstraktionen und Relationen ein Erkennen und Kommunizieren gleicher Konzepte und dadurch eine disziplinunabhängige semantische Vernetzung der Informationen. Durch die semantische Modellierung in Form von sog. Pfaden ist eine nachhaltige Interpretierbarkeit der Zusammenhänge von unterschiedlichen Informationen möglich, die für die inhaltliche Erschließung der Quellenbestände von Bedeutung ist. Die Pfade wiederum sind netzwerkartig miteinander verbunden. So kann z. B. nachvollzogen werden, in welchem Verhältnis eine Person zu einem Friedensereignis oder zu einem Objekt steht, beides kann für die Forschungsfragen nach Funktion des jeweiligen Quelleninhaltes von Interesse sein.

Für die Veröffentlichung der Ergebnisse in einem virtuellen Themenportal können zur besseren Strukturierung und auch um Abhängigkeiten darzustellen, Informationen hierarchisch in Beziehung gesetzt werden, wie Friedensschlüsse und auf ihnen basierende Anlässe oder Allegorien zu übergeordneten Bildtopoi. Auf allen hierarchischen Stufen bleiben die entsprechenden Eigenschaften und Relationen der entsprechenden Klassen erhalten und können demzufolge immer mit abgebildet und abgefragt werden.

Die unterschiedlichen Informationen werden in spezifisch modellierten Masken erfasst, in deren Feldern Normdaten und Vokabulare hinterlegt sind. Durch verschiedene systemimmanente Eigenschaften können Wissenschaftler sehr schnell in einer Objektmaske zugehörige Dokumente und Objekte angezeigt bekommen. Für den Benutzer dient dies bei ca. 2000 angestrebten Einträgen der Übersichtlichkeit, so dass auch auf dieser Ebene die Vernetzung sichtbar sein wird.

3.2 Anwendungsbeispiel Projekt „MUSICES“

WissKI dient dem Projekt als Datenbank für die zu untersuchenden Musikinstrumente und als Kommunikationsplattform. Darüber hinaus ist das System in der Lage, den kompletten Untersuchungsablauf sowie die Messergebnisse und die erzeugten 3D-Daten jedes einzelnen Objekts, zusammengefasst das im Standard enthaltene Dokumentationsschema und das Netzwerk der Metadaten, abzubilden. Die zu

erfassenden Metadaten beinhalten nicht nur die objektbezogenen des kulturwissenschaftlichen Bereichs, sondern auch die vom Fraunhofer Institut zu dokumentierenden Messparameter, wie die Röntgenspannung, die applizierte Strahlungsdosis, aber auch Informationen zu den CT-Anlagen. Für die Erfassung der Projektdaten in einem gemeinsamen Kontext wurde eine Anwendungsontologie, basierend auf dem CIDOC CRM entwickelt, die ebenfalls Teil des im Projekt zu entwickelnden Standards ist. Durch eine klare Definition der Metadaten, die sich auch in der Modellierung der Pfadstrukturen niederschlägt, entsteht eine Datenstruktur, die eine weitere Nutzbarkeit und Interpretierbarkeit der Projektergebnisse gewährleistet.

Durch die Verwendung des CIDOC CRM können die Metadaten in das museumsinterne Objektdokumentationssystem und darüber hinaus in internationale Portale integriert werden. Im Rahmen des EU-Projekts MIMO [7] konnte mit MIMO-LIDO ein Metadatenmodell für Musikinstrumente entwickelt werden, das die Grunddatenerfassung und die Zuordnung zu Sammlungskontexten standardisiert. Das Metadatenmodell für die 3D-CT-Aufnahmen des MUSICES-Projekts wird in MIMO-LIDO integriert, steht darüber hinaus aber auch als eigenständige Domänenontologie zur Verfügung. Für den Bereich der Erforschung von Musikinstrumenten und ihrer künftigen Erfassung, insbesondere im Hinblick auf 3D-CT-Maßnahmen, wird das MUSICES-Projekt Wegbereiter für einen Standard sein, der auf verschiedenen bestehenden Standards des kulturellen Bereichs aufbaut und diese für einen spezifischen Anwendungsfall ergänzt. Durch die Publikation mit WissKI und internationalen Portalen kann garantiert werden, dass die Projektdaten verfügbar und zitierbar sind.

In beiden Forschungsprojekten, obgleich ihrer unterschiedlichen Disziplinen und Objektgattungen, kann durch Anwendung des CIDOC CRM eine nachhaltige Interpretierbarkeit und Austauschbarkeit der in den Projekten erhobenen Daten am GNM gewährleistet werden. In Verbindung mit WissKI sind alle anwendungsspezifischen Anforderungen abgedeckt. Durch seine Systemarchitektur ist WissKI flexibel genug, auch auf sich während der Projektlaufzeit neu ergebende Forschungsfragen zu reagieren.

Fußnoten

1. MUSICES: Sebastian Kirsch1, Frank Bär1, Theobald Fuchs2, Christian Kretzer2, Markus Raquet1, Gabriele Scholz2, Rebecca Wagner2, Meike Wolters-Rosbach1; 1 Germanisches Nationalmuseum, Nürnberg; 2 Fraunhofer-Entwicklungszentrum Röntgentechnik EZRT, Fürth
2. Das Leibniz-Institut für Europäische Geschichte, Mainz, untersucht Friedenspredigten, die Herzog August Bibliothek, Wolfenbüttel, Dichtungen und Festschriften, das Germanische Nationalmuseum Objekte aus den graphischen und numismatischen Sammlungen, das Deutsche Historische Institut, Rom, Kantaten, Oratorien und Festmusiken vor allem in Bezug auf Italien und das Tadeusz Manteuffel Institut für Geschichte der Polnischen Akademie der Wissenschaften, Warschau, die Friedensrepräsentationen in den östlichen Gebieten Europas.
3. Diese Ontologie wurde vom International Committee for Documentation (CIDOC) als Teil des International Council of Museums (ICOM) erstellt (URL: http://www.cidoc-crm.org/), wobei das Germanische Nationalmuseum federführend beteiligt war.
4. URL: http://erlangen-crm.org / (25.08.2016).
5. OWL= Web Ontology Language, vgl. URL: https://www.w3.org/TR/owl2-overview/ (25.08.2016).
6. WissKI = Wissenschaftliche Kommunikations-Infrastruktur, URL: http://wiss-ki.eu /) basierend auf dem Open-Source Content Management System Drupal (URL: http://drupal.org /), und wurde in Zusammenarbeit zwischen dem Germanischen Nationalmuseum, Nürnberg, dem Zoologischen Forschungsmuseum Alexander Koenig, Bonn und der Friedrich-Alexander-Universität Erlangen-Nürnberg entwickelt.
7. Musical Instrument Museums Online (URL: www.mimo-international.com .). Während der Projektlaufzeit 2009 bis 2011 wurden rund 50.000 Musikinstrumente in öffentlichen Sammlungen digitalisiert und über MIMO-DB zugänglich gemacht (URL: http://www.mimo-db.eu/ (25.8.2016).

Bibliographie

Doerr, Martin / Lampe, Karl-Heinz / Krause, Siegfried (2011): *Definition des CIDOC Conceptual Reference Model Version 5.0.1*. autor. durch die CIDOC CRM Special Interest Group (SIG)

(= Beiträge zur Museologie 1). Berlin: ICOM Deutschland.

Görz, Günther (2011): „WissKI: Semantische Annotation, Wissensverarbeitung und Wissenschaftskommunikation in einer virtuellen Forschungsumgebung“, in: *Kunstgeschichte. Open Peer Reviewed Journal* urn:nbn:de:bvb:355-kuge-167-7 [letzter Zugriff 22. November 2016].

Hohmann, Georg / Fichtner, Mark (2015): „Chancen und Herausforderungen in der praktischen Anwendung von Ontologien für das Kulturerbe“, in: Robertson – von Trotta, Caroline Y. / Schneider, Ralf Y. (eds.): *Digitales Kulturerbe*. Bewahrung und Zugänglichkeit in der wissenschaftlichen Praxis. (= Kulturelle Überlieferung – digital 2). Karlsruhe: KIT Scientific Publishing 115-128.

Stein, Regine / Gottschewski, Jürgen / Heuchert, Regina / Ermert, Axel / Hagedorn-Saupe, Monika / Hansen, Hans-Jürgen / Saro, Carlos / Scheffel, Regine / Schulte-Dornberg, Gisela (2005): *Das CIDOC Conceptual Reference Model. Eine Hilfe für den Datenaustausch?* (= Mitteilungen und berichte aus dem Institut für Museumskunde 31). Berlin: Institut für Museumskunde.

Nachhaltige Erschließung umfangreicher handschriftlicher Überlieferungen. Ein Fallbeispiel

Faßhauer, Vera
fasshauer@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Angesichts stetig wachsender Kapazitäten zur Speicherung großer Datenmengen nutzen Bibliotheken und Archive zunehmend die Möglichkeit, ihre Sammlungen zu digitalisieren und die Faksimiles online bereitzustellen. Rein konservatorischen Erwägungen folgend, belassen sie es dabei häufig bei der Erfassung der Metadaten und verzichten auf die weiterreichende inhaltliche Erschließung des Materials. So bleibt es oftmals allein dem Nutzer überlassen, sich einen Zugang zu den Inhalten der Sammlungen zu verschaffen.

Sofern es sich dabei um Druckwerke handelt, ist dieses Vorgehen durchaus hinreichend, zumal die Fähigkeit zur Lektüre von Antiqua- und Frakturdrucken zumindest im deutschsprachigen Raum allgemein vorausgesetzt werden kann. Da mit Hilfe der OCR-Technologie inzwischen selbst bei der automatischen Erkennung der Frakturschrift sehr gute Ergebnisse erzielt werden können, werden digitalisierte Druckwerke auch jenseits der genauen inhaltlichen Erfassung nutzbar, indem sie durch *distant reading* und statistische Zugänge erschlossen werden können.

Anders verhält es sich bei historischen Handschriften: Da heutzutage nur sehr wenige Personen über hinreichende paläographische Kenntnisse verfügen, stellen digitale Reproduktionen handgeschriebener Dokumente für den größten Teil des Publikums nicht viel mehr als bloße Abbildungen historischer Artefakte dar, die in ihrer Materialität zwar eine ganz bestimmte Oberflächenstruktur aufweisen, aber die darin transportierten Inhalte nur wenigen erfahrenen Lesern preisgeben. Zusätzlich erschwert wird die Lektüre im Fall von Tagebuchaufzeichnungen oder Notizbüchern, die nur selten auch für fremde Augen bestimmt waren.

Die Gewährleistung eines unbeschränkten und langfristigen Zugangs zu digitalisierten historischen Handschriftenarchiven ist also nicht *per se* gleichbedeutend mit einer unbegrenzten Zugänglichkeit, Nutzbarkeit und Weiterverwertbarkeit ihrer Inhalte. Eine wichtige Aufgabe der *Digital Humanities* bei der nachhaltigen Pflege des kulturellen Erbes ist deshalb eine über die bloß konservierende Ablichtung hinausgehende Erschließung der in diesen Textbeständen enthaltenen Informationen. Die Fragestellung ist also: Wie lassen sich diese Daten erfassen und für *close*- wie auch für *distant reading-Prozesse* aufbereiten? Lässt sich ein Zugang schaffen, ohne den gesamten Bestand manuell zu bearbeiten? Und inwieweit kann die klassische paläographische Hand- und Kopfarbeit durch automatisierte Prozesse ersetzt werden? Der Beitrag stellt diese Problemlage zunächst am Fallbeispiel der Senckenberg-Tagebücher exemplarisch dar und zeigt anschließend eine Lösungsstrategie auf, bei der manuelle und digitale Methoden kombiniert zum Einsatz kommen und bereits vorhandene, frei zugängliche Software verwendet wird.

Der Frankfurter Arzt Johann Christian Senckenberg (1707–1772) hinterließ handschriftliche Aufzeichnungen im Umfang von 53 Quartbänden mit je etwa 700 Seiten.

Während die späteren Bände einesteils in ausführlichen ärztlichen Fallstudien und anderenteils in kritischen Bemerkungen über die sittlichen Missstände der Reichsstadt bestehen, befassen sich die mit *Observationes in me ipso factae* übertitelten ersten dreizehn Jahrgänge hauptsächlich mit dem Schreiber selbst. Da Senckenberg dem radikalen Pietismus nahestand und sich ganz aus dem kirchlichen Gemeindeleben zurückgezogen hatte, erfüllten die frühen Tagebücher hauptsächlich die Funktion eines religiösen Gewissensspiegels. Darüber hinaus notierte er über Jahrzehnte hinweg täglich seinen Speiseplan, sein Bewegungspensum und seine Stoffwechselaktivität ebenso detailliert wie die jeweilige Wetterlage, die Umgebungstemperatur und den Luftdruck, die er mit den wechselnden Zustände seines Gemüts und mit äußeren Umwelteinflüssen in Beziehung setzte. Der Zweck dieser akribischen Beobachtungen war seine diätetische und moralische Selbstoptimierung, welche sowohl eine untadelige Lebensführung im Diesseits als auch seine Erlösung im Jenseits gewährleisten sollte. Zugleich dienten sie der Erfassung und Deutung von Korrelationen zwischen Vorgängen in Leib, Seele, Natur und Kosmos.

Diese Aufzeichnungen stellen sich nicht nur dem heutigen Publikum als Big Data dar, sondern wurden bereits von ihrem Autor als riesiger Datenpool konzipiert: Zeitweise brachte er täglich bis zu 5000 Wörter in deutscher und lateinischer Sprache zu Papier, so dass er in manchen der insgesamt 43 Jahrgänge ca. 2600 Seiten sehr eng mit jeweils etwa 900 Wörtern beschrieb. Zugleich pietistisches Selbstzeugnis und wissenschaftliche Aufzeichnungsform, ist dieser schriftlich fixierte und weltweit einzigartige Erfahrungsschatz eine Fundgrube für die Erforschung der frühneuzeitlichen Religions- und Wissenschaftsgeschichte. Darüber hinaus wirft er neue historische Schlaglichter auf die aktuell diskutierten Möglichkeiten und Grenzen der Nutzung großer Datensammlungen und ihr Verhältnis zur Theorie (vgl. Anderson 2008; boyd et al. 2012, Rosenberg 2014).

Mit Förderung durch die Dr. Senckenbergische Stiftung wurden die insgesamt ca. 40.000 Quartseiten in hochaufgelöster Form digitalisiert und von der Universitätsbibliothek Frankfurt unter Open Access-Bedingungen online zur Verfügung gestellt (UB Frankfurt 2013–2016). Am Frankfurter Institut für Deutsche Literatur und ihre Didaktik entsteht derzeit eine TEI/XML-basierte Online-Edition der Aufzeichnungen, welche gleichfalls von der durch den Autor selbst begründeten Stiftung finanziert wird. In Anbetracht ihres riesigen Umfangs und der schwer entzifferbaren Handschrift Senckenbergs ist eine zeitnah fertigstellbare Volltextedition des Gesamtbestandes schwerlich möglich und wäre aufgrund der bei dieser Aufzeichnungspraxis naturgemäß häufig auftretenden inhaltlichen Redundanzen auch nicht sinnvoll. Aus diesem Grund wurde im Vorfeld eine repräsentative Bandauswahl getroffen, welche nach den Maßgaben der historischen Signifikanz, der thematischen Vielfalt und der größtmöglichen Vermeidung von Redundanzen erfolgte. Die inhaltliche Komplexität und die auf schnelle Erfassung großer Datenmengen ausgerichtete Schreibroutine des Autors machen zudem eine Transkriptionsweise erforderlich, die weit über die diplomatisch-zeichengetreue Textwiedergabe hinausgeht: Abgesehen von der Tatsache, dass es sich um einen halb frühneuhochdeutschen und halb lateinischen Text handelt und der Schreiber oftmals mehrfach in einem Satz zwischen beiden Sprachen hin- und herwechselt, sind viele der Sätze so komplex, dass der Leser zum Verständnis auf alle verfügbaren grammatischen Merkmale angewiesen ist. Vor allem die morphologischen Merkmale sind aber im Deutschen wie auch im Lateinischen hauptsächlich in eben jenen Wortendungen enthalten, welche häufig durch Abkürzung entfallen. Ein ähnliches Problem besteht auch hinsichtlich der Symbole, die größtenteils dem alchemistischen Kontext entstammen: Sie können ein ganzes Wort oder auch nur einen Teil davon ersetzen, bis zu vier verschiedene Wortbedeutungen und noch viel mehr grammatische Formen repräsentieren und in völlig verschiedenen semantischen Umgebungen erscheinen. Um dem Leser einen hinreichenden Zugang zum Sprachgebrauch des Autors zu bieten und ein Textverständnis überhaupt erst zu ermöglichen, müssen Abkürzungen und Symbole ihrem kontextspezifischen Zusammenhang entsprechend aufgelöst und sowohl semantisch als auch grammatikalisch in den Text eingepasst werden.

Auf den ersten Blick scheinen digitale Methoden hier kaum weiterzuhelfen: Zu wenig deutlich ist die Schrift, zu komplex die Inhalte, zu spezifisch das Vokabular und zu mehrdeutig die einzelnen Zeichen. Hinzu kommt noch, dass sich sowohl Senckenbergs Handschrift als auch die Inhalte seiner Aufzeichnungen im Verlauf von vier Jahrzehnten stark veränderten und mithin ganz neue graphische Muster hervorbrachten. Wenngleich die Transkription der Texte nur händisch erfolgen kann, wird dadurch doch

ihre Maschinenlesbarkeit überhaupt erst gewährleistet und damit die grundlegende Voraussetzung für automatisierte Prozesse sowie die Anwendung, Weiterentwicklung und Schulung der sie ermöglichenden Technologien geschaffen. So erfordert das Training des Tools Transkribus (Universität Innsbruck o.J.) zunächst einmal eine ausreichende Menge an manuell erzeugten und präzisen Texttranskriptionen und die anschließende händische Überarbeitung des Outputs (vgl. Transkribus Wiki o.J.). Aufgrund der wachsenden Nachlässigkeit der Handschrift und der inhaltlichen Heterogenität der drei Unterbestände muss der Lernprozess für jeden Teilbestand separat erfolgen. Nach Abschluss dieses Lernprozesses ist jedoch zumindest eine halbautomatische Texterfassung möglich. Der erkannte Text kann anschließend elektronisch durchsucht und wissenschaftlich ausgewertet werden.

Ein ähnliches Verhältnis zwischen manuellen und automatisierten Prozessen besteht hinsichtlich der inhaltlichen Erschließung der Texte. Da sie von einem einzigen Schreiber mit umfassender grammatischer Bildung stammen, liegt nur eine geringe orthographische Varianz bei der Schreibung ein- und desselben Wortes vor. Anders als in heterogenen Korpora, die Texte mehrerer Schreiber mit unterschiedlichem Bildungshintergrund und sprachgeografischer Herkunft versammeln, ist deshalb eine vorherige händische Normierung der Grafie nicht notwendig (vgl. demgegenüber Faßhauer et al. 2013, Faßhauer et al. 2014). Mit Hilfe der vorliegenden Transkriptionen kann deshalb ein effizientes Training des Tools TreeTagger (Schmid 1994-) für das Frühneuhochdeutsche und Neulateinische vorgenommen werden. Die halbautomatisch generierten Lemmata und Part-of-Speech-Tags, welche sowohl für die manuellen Transkriptionen als auch für die automatisch erfassten Texte erstellt wurden, werden anschließend in den Partitur-Editor der Software EXMARaLDA (Hedeland et al. o.J.) eingespielt. Mit dem zugehörigen Analysetool EXAKT werden per RegEx-Suche auf der Lemmaspur zunächst alle Nomina herausgefiltert und in einem manuellen Prozess Schlagwörter ausgewählt (ähnlich auch Biehl et al. 2015). Aus der Untermenge der großgeschriebenen Substantive, die sich mittels der automatischen Sortierfunktion der Trefferliste leicht ermitteln lassen, werden alle Personen- und Ortsnamen entnommen. Anhand dieser Recherchezugänge kann nun das gesamte Korpus systematisch recherchiert werden. Die von EXAKT angebotenen Anfragen über RegEx und Levenshtein-Distanzen ermöglichen dabei eine schreibweisentolerante Begriffsermittlung, wodurch mancher HTR-Lesefehler überwunden werden kann.

Bibliographie

Anderson, Chris (2008): „The End of Theory: The Data Deluge Makes the Scientific Method Obsolete“, in: *Wired Magazine* http://www.wired.com/2008/06/pb-theory/

Biehl, Theresia / Lorenz, Anne / Osierenski, Dirk (2015): „Exilnetz33. Ein Forschungsportal als Such- und Visualisierungsinstrument“, in: Baum, Constanze / Stäcker, Thomas (eds.): *Grenzen und Möglichkeiten der Digital Humanities* (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1).

Boyd, Danah / Crawford, Kate (2012): „Critical Questions for Big Data“, in: *Information, Communication & Society* 15 (5): 662–679 10.1080/1369118X.2012.678878.

Fasshauer, Vera / Lühr, Rosemarie / Prutscher, Daniela /Seidel, Henry (2013): Dokumentation der Annotationsrichtlinien für das Korpus *Frühneuzeitliche Fürstinnenkorrespondenzen im mitteldeutschen Raum*. dwee.eu/Rosemarie_Luehr/userfiles/downloads/Projekte/Dokumentation.pdf.

Fasshauer, Vera / Lühr, Rosemarie / Prutscher, Daniela /Seidel, Henry (2014): *Fürstinnenkorrespondenz* (version 1.1), Universität Jena, DFG. LAUDATIO Repository. http://www.indogermanistik.uni-jena.de/Web/Projekte/Fuerstinnenkorr.htm .

Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (o.J.): *EXMARaLDA. Werkzeuge für mündliche Korpora* http://www.exmaralda.org/ [letzter Zugriff 21. März 2016].

Rosenberg, Daniel (2014): „Daten vor Fakten“, in: Reichert, Ramón (ed.): *Big Data: Analysen zum digitalen Wandel von Wissen, Macht und Ökonomie*. Bielefeld: transcript Verlag 133–156.

Schmid, Helmut (1994-): *TreeTagger. A part-of-speech tagger for many languages*. http://www.cis.unimuenchen.de/~schmid/tools/TreeTagger/ [letzter Zugriff 20. August 2016].

Transkribus Wiki (o.J.): *Transkribus-Benutzeranleitung*. https://transkribus.eu/wikiDe/index.php/Hauptseite [letzter Zugriff 20. August 2016].

UB Frankfurt =Universitätsbibliothek Frankfurt am Main (2013-2016): *Nachlass Johann Christian Senckenberg*. http://sammlungen.ub.uni-frankfurt.de/senckenberg/nav/index/all [letzter Zugriff 20. August 2016].

Universität Innsbruck (o.J.): *Transkribus*. https://transkribus.eu/Transkribus/ [letzter Zugriff 20. August 2016].

Nachhaltige Konzeptionsmethoden für Digital Humanities Projekte am Beispiel der Goethe-PROPYLÄEN

Kasper, Dominik
dominik.kasper@adwmainz.de
Akademie der Wissenschaften und der Literatur, Deutschland

Grüntgens, Max
max.gruentgens@adwmainz.de
Akademie der Wissenschaften und der Literatur, Deutschland

Vor dem Hintergrund des Akademievorhabens *PROPYLÄEN. Forschungsplattform zu Goethes Biographica* werden zwei aktuelle Konzeptionsmethoden aus dem Bereich des Software-Engineering vorgestellt, die sich unserer Erfahrung nach als geeignete Grundlage für nachhaltige Konzeptionsprozesse in Digital Humanities Projekten erwiesen haben: *Domain Driven Design* sowie *Behaviour Driven Development*. Zunächst erfolgt ein Überblick über Aspekte digitaler Nachhaltigkeit bei der Beantragung geisteswissenschaftlicher Langzeitvorhaben im Akademienprogramm.

Die PROPYLÄEN sindein 2015 im Akademienprogramm gestartetes Forschungsvorhaben der Klassik Stiftung Weimar, der Sächsischen Akademie der Wissenschaften zu Leipzig und der Akademie der Wissenschaften und der Literatur Mainz (http://www.goethe-biographica.de). Geplant ist bei einer Gesamtlaufzeit von 25 Jahren Fortführung und Abschluss von vier (Print-)Publikationsreihen: Die Editionen der Briefe von sowie an Goethe, seine Tagebücher und die Zeugnisse seiner Begegnungen und Gespräche. Die gleichzeitig entstehende Forschungsplattform wird sukzessive alle Biographica in einer technisch und inhaltlich offenen Infrastruktur bereitstellen sowie Recherchezugänge und Visualisierungen auf den Datenbestand anbieten.

Eine Voraussetzung für die Aufnahme eines Neuvorhabens in das Akademienprogramm ist ein ausgearbeitetes Digitalisierungskonzept, zu dem unter anderem eine Strategie zur Langzeitarchivierung und Langzeitverfügbarkeit der Forschungsergebnisse gehört (Vgl. Herrmann 2016: 9). Häufig konzentriert sich dieses auf eine enge Definition von Forschungsdaten, die unter diesem Begriff lediglich edierten Text versteht. Im Falle einer Forschungsplattform wie der PROPYLÄEN sind nicht nur die in der Datenschicht – Faksimiles, Editionstext, kurz die Gesamtheit der digitalen Forschungsdaten – befindlichen Komponenten für eine nachhaltige und langfristige Bereitstellung vorzusehen (zu Archivierungsschicht und Präsentationsschicht vgl. Pempe 2012: 141). Daneben sollte auch die Auswahl geeigneter Technologien ein Teil der Antragsstrategie sein. Unter Technologie sei alles vom Datenformat über ein Content-Management-System bis zur virtuellen Forschungsumgebung verstanden. Angesichts des Verhältnisses zwischen langfristiger Projektförderung – die Akademienvorhaben werden 12 bis 25 Jahre gefördert – und der technischen Entwicklung ist die Vorstellung der einmaligen Einrichtung einer digitalen Arbeitsumgebung und Publikationsplattform am Anfang der Projektlaufzeit, welche daraufhin für 25 Jahre verwendet würde, nur mit begleitender Wartung und Weiterentwicklung denkbar. Wird das gewählte Datenformat in 30 Jahren noch maschinenlesbar sein? Absolute Sicherheit gibt es in diesem Fall nicht, aber allgemein gilt die Verwendung von (semistrukturierten) Rein-Textformaten – wie XML nach TEI – gegenüber proprietären Formaten als nachhaltiger.

Der *Mehrwert der elektronischen Fassung* entsteht aber vornehmlich durch neue Verbindungs-, Gruppier-, Sortier-, Filter- und Suchmöglichkeiten, die auf dynamischen Verarbeitungsmechanismen, also auf der Geschäftslogik der Anwendung, basieren. Diese nicht-statischen Elemente, die sich insbesondere in der Präsentationsschicht einer geisteswissenschaftlichen Webanwendung konkretisieren, müssen in ihrer Funktionalität genauso nachvollziehbar und reproduzierbar über das Projektende hinaus zur Verfügung stehen.

Komplexer als die Datenschicht ist demnach die dauerhafte Erhaltung der Applikationslogik und Präsentationsschicht: Wie kann die webbasierte Verarbeitung von Anfragen und die Wiedergabe von Ergebnissen auch 10 Jahre nach Projektende noch funktionsfähig

gehalten werden? Eine aktuell diskutierte Strategie ist die virtuelle Kapselung sämtlicher Softwarekomponenten einer Applikation, die künftig ein Emulieren aller Komponenten auf aktuellen Betriebssystemen ermöglicht (siehe dazu bspw. http://recomputation.org/). Vorhaben wie die PROPYLÄEN müssen dieser Herausforderung während der Projektlaufzeit begegnen, indem sie eine hohe Flexibilität sowie stete Aktualisierung und Anpassung der Technologien ihrer Infrastruktur in der Grundplanung vorsehen. Um diesen Prozess auf eine transparente Grundlage zu stellen, gilt es bei der informationstechnischen Gestaltung des Forschungsvorhabens eine möglichst *gegenstandsnahe* Abstraktion zu wählen und die Anforderungen an die Software auf *formalisierte* Weise zu dokumentieren. Folgende Annahmen und Schlüsse gehen dieser Überlegung voran:

- Es wird während der Projektlaufzeit intern wie extern neue Erkenntnisse zu Goethes Zeit und Wirken geben, die unseren Blick auf den Forschungsgegenstand verändern. Dennoch ist es möglich, übergreifende Wissensobjekte zu identifizieren und in Form von informatischen *Entitäten* abzubilden.
- *Technologieunabhängige, anwendungsorientierte* und *allgemeinverständliche* Funktionsbeschreibungen der Forschungsplattform sind formulierbar. Neue Erkenntnisse im Bereich der Goethe-Forschung wie im Bereich der *usability* werden diese Anforderungen innerhalb der Projektlaufzeit höchstwahrscheinlich verändern.

In der *Konzeptionsphase* einer nachhaltigen digitalen Grundlage für den Daten-, Präsentations- und Applikationsteil der Forschungsplattform PROPYLÄEN sind daher zwei Fragen leitend, deren Beantwortung uns nicht nur für Langzeitvorhaben mit Digital-Humanities-Anteil zentral erscheint (vgl. zu Prozessphasen und zur Konzeption hier und folgend: Schrade 2016: Step 14–21, zur Konzeption Step 17):

- Wie lassen sich *Biographica* Goethes für technische und geisteswissenschaftliche Anforderungen nachhaltig modellieren bei gleichzeitiger Bewahrung der Flexibilität in der Anwendungsarchitektur für die Integration gegebenenfalls noch nicht bekannter digitaler Ressourcen?
- Wie lassen sich die funktionalen Anforderungen an eine digitale Rechercheumgebung aus der Fachcommunity nachhaltig formulieren und dokumentieren?

Zur Beantwortung können zwei Konzeptionsmethoden aus dem Bereich des Software-Engineering fruchtbar gemacht werden: *Domain Driven Design (DDD)* (siehe Evans 2003, Vernon 2013) und *Behavior Driven Development (BDD)* (siehe North 2016) . Beiden ist gemein, dass sie unabhängig von Datenformaten, Programmiersprachen oder Präsentationstechnologien operieren.

DDD nimmt an, dass einem Konzeptions- bzw. Entwicklungsprozess dann die größten Erfolgschancen und daraus resultierend die beste Nachhaltigkeit zukommen, wenn die *virtuellen* Komponenten des informatischen Modells ausgehend von ihren *realen* Entsprechungen gebildet werden. Damit kommt der fachwissenschaftlichen Logik, die sich aus dem Wissen von Domänen-Experten (im Anwendungsfall Goethe-Philologen) und deren Niederlegung in publizierten Editionsbänden ableitet, eine zentrale Rolle für die Modellierung zu. Die Ausmodellierung der Wissensdomäne entsteht iterativ in einem konstanten kommunikativen Prozess aller Projektbeteiligten. *DDD* löst die gängige und nicht optimale Praxis einer ausschließlichen Bedarfsformulierung durch Geisteswissenschaftler und einer darauffolgenden Umsetzung durch Informatiker zugunsten eines gemeinsamen Modellierungsprozesses unter Verwendung einer für alle Beteiligten verständlichen (*ubiquitären*) Sprache auf. Dieser im Rahmen eines Projektes von allen Beteiligten zu entwickelnden Sprache liegen eine Reihe von Komponenten für ein Modell der Wissensdomäne (*domain model*) zugrunde, das in seiner Gesamtheit alle Eigenschaften, Beziehungen und „Geschäftsprozessen“ der zukünftigen Anwendung abbilden kann. Die wichtigsten Komponenten für die Fachdomäne „PROPYLÄEN“ sind:

- Objekte mit eigener Identität (*entities)*: Briefe, Faksimiles, Personen, Orte, Werke
- Objekte, die sich über die Gesamtheit ihrer Eigenschaften definieren (*value objects*): Geokoordinaten (Länge-Breite), Datierungen (Anfangs-Enddatum), Korrespondenzvorgang (Sender-Empfänger) etc.

- Die Beziehungen dieser Objekte untereinander (*associations*): Person → erwähnt in: Brief
- Zusammenfassungen von *entities, value objects* und *associations* zu Aggregaten (*aggregates*); der Zugriff auf einzelne Komponenten eines Aggregates darf ausschließlich über die Entität an der Aggregat-Wurzel (*aggregate root*) erfolgen. Aggregate gewährleisen logische Konsistenz und funktionale Integrität einer Anwendungs- bzw. Wissensdomäne: Brief (Brieftext, Kommentar, Faksimile, Datierung, im Brief erwähnte Orte und Personen)
- Dienste zwischen Objekten der Wissensdomäne (*services*): Sortiere nach Datum etc.

Die Fachdomäne kann grafisch dargestellt werden. Abbildung 1 dient der Illustration des DDD-Konzepts am Beispiel eines Teils der PROPYLÄEN-Domäne.

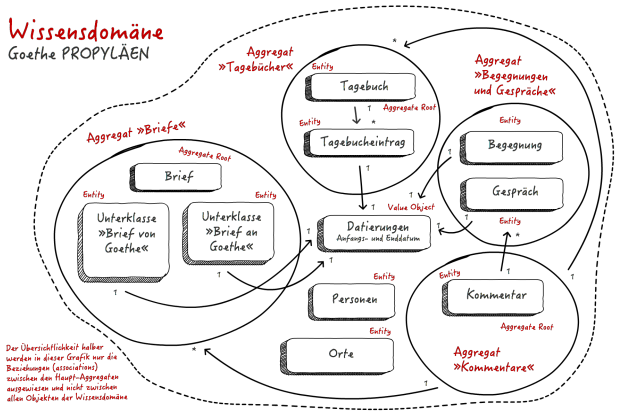


Abbildung 1: Wissensdomäne Goethe PROPYLÄEN

Dan Norths Idee des *BDD* ist, Anforderungen an die geplante Applikation in Form von *user stories* und verhaltensorientierten Testszenarien im Vorfeld der Realisierung unter Zuhilfenahme einer ubiquitären Sprache zu beschreiben. Durch die Verwendung von Schlüsselwörtern in formalisierter Notation lässt sich eine einheitliche und leicht verständliche Dokumentation erstellen. Aufgrund starker Nutzerorientierung wird *BDD* im Qualitätsmanagement von Softwareanwendungen zu den *Akzeptanztests* gerechnet (neben Unit-Tests, Funktionstests, Performance-Tests oder manuellen Tests).

Abbildung 2 zeigt ein BDD-Testszenario in den Goethe-PROPYLÄEN, das auf Gherkin-Notation basiert (Siehe https://github.com/cucumber/cucumber/wiki/Gherkin , Stand 26.08.2016). Die Schlüsselwörter *Funktionalität, Beschreibung und Details* enthalten übergeordnete Informationen, während im Bereich der *Szenarios* mit Hilfe der Operatoren *Szenario bzw. Gege ben, Wenn, Dann, Und* einzelne Anforderungen formuliert werden. Die Ausführungen können dabei allgemein gehalten werden und lassen sich im späteren Verlauf auffächern.

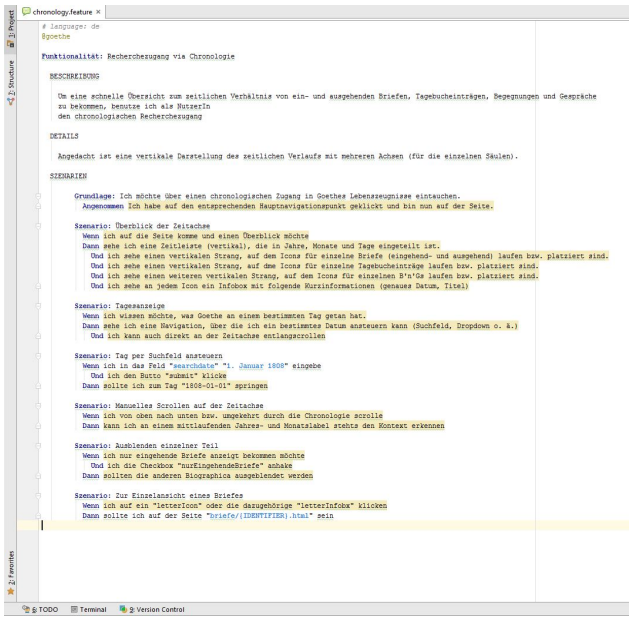


Abbildung 2: BDD-Testszenario in Gherkin-Notation

Ein Vorteil ist, dass Szenarios durch den Einsatz einer verständlichen, formalisierten Sprache nicht nur gemeinsam zwischen Geisteswissenschaftlern und Informatikern während des Entwicklungsprozess formuliert werden, sondern dass bei syntaktisch korrekter Formulierung diese als Grundlage für automatisierte Testverfahren der Präsentationsschicht einer Webanwendung herangezogen werden können (bspw. mit dem BDD Testing-Framework behat; siehe http://docs.behat.org (Stand 26.08.2016)). BDD ermöglicht somit eine zeitökonomische Überprüfung aller Bereiche und Funktionalitäten der Präsentationsschicht bei gleichzeitiger Dokumentation der Funktionalitäten in Reintext und natürlicher Sprache. Änderungen in der technischen Infrastruktur einer Plattform (Softwarewechsel der Datenbankschicht, Einführung neuer Programmroutinen) können mit deutlich gesteigerter Verlässlichkeit im Hinblick auf eine gleichbleibende, dauerhafte Funktion der Webanwendung durchgeführt werden.

Die gemeinsame Modellierung einer Wissensdomäne mittels DDD und BDD führt notwendigerweise zu intensiver Auseinandersetzung mit allen geistes- wie informationstechnischen Aspekten des Projektes. Dadurch wird projektintern eine kommunikative Ebene entstehen, die eine gemeinsame und unmißverständliche Sprache verwendet und perspektivisch eine verbesserte Tiefenschärfe in Bezug auf funktionale Aspekte der zu realisierenden Forschungsanwendung ermöglicht.

DDD und BDD helfen als Konzeptionsmethoden, alle gemeinsam getroffenen Entscheidungen über die gesamte Projektlaufzeit transparent und nachvollziehbar zu dokumentieren. Da es sich gleichzeitig im Bezug auf die BDD-Akzeptanztests um „ausführbares Wissen“ handelt, wird eine dauerhafte und gleichbleibende Funktionalität der Präsentationsschicht geisteswissenschaftlicher Webanwendungen gewährleistet. Dies trägt erheblich zu gesteigerter Nachhaltigkeit digitaler Komponenten eines geisteswissenschaftlichen Forschungsprojektes bei.

Bibliographie

Evans, Eric (2003): *Domain-Driven Design*. Tackling Complexity in the Heart of Software. Boston et al.

Herrmann, Dieter (2016): „E-Humanties im Akademienprogramm“, in: Union der Deutschen Akademien der Wissenschaften (Hrsg.): *Die Wissenschaftsakademien – Wissensspeicher für die Zukunft. Forschungsprojekte im Akademienprogramm*. Berlin / Mainz 9–11.

North, Dan (2016): *Introducing BDD*. https://dannorth.net/introducing-bdd/ [letzter Zugriff 26. August 2016].

Pempe, Wolfgang (2012): „Geisteswissenschaften“, in: Neuroth, Heike et al. (eds.): *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*. Göttingen 137–159.

Schrade, Torsten (2016): *Nachhaltige Online-Applikationen in den Geisteswissenschaften – Modellierung und Implementierung*. Vortrag vom 11. April 2016, Hochschule Mainz, http://metacontext.github.io/nachhaltige-online-apps/ [letzter Zugriff 26. August 2016].

Vernon, Vaugh (2013): *Implementing Domain-Driven Design*. Boston et al.

Nachhaltige Softwareentwicklung in den Digital Humanities. Konzepte und Methoden.

Schrade, Torsten
Torsten.Schrade@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland

Ausgehend von den umfangreichen Infrastrukturinitiativen der vergangenen Jahre existieren inzwischen vielfältige digitale Ressourcen, Werkzeuge und Dienste, die von einer lebhaften digitalen Forschungskultur in den Geisteswissenschaften zeugen. Arbeitsgruppen wie beispielsweise NESTOR oder auch die DINI-Initiative haben es sich zur Aufgabe gemacht, Empfehlungen und *best practices* für den gesamten Lebenszyklus digitaler geisteswissenschaftlicher Forschungsprojekte (Datenerfassung, Datenverwaltung, Datenpublikation, Datenarchivierung) zu entwickeln. Mit dem von DARIAH und TextGrid initiierten Memorandum zur nachhaltigen Bereitstellung digitaler Forschungsinfrastrukturen für die Geistes- und Kulturwissenschaften in Deutschland ist das Thema ‚Digitale Nachhaltigkeit‘ ganz besonders in den Fokus gerückt (s. http://dhd-blog.org/?p=6559).

Während das Bewußtsein für eine nachhaltige Erschließung kultureller Objekte durch den Einsatz entsprechender Datenformate und -standards inzwischen als hoch eingeschätzt werden kann, gelingt eine nachhaltige Integration von Softwarewerkzeugen in die konkrete Forschungswirklichkeit geisteswissenschaftlicher Projekte noch nicht immer. Die Gründe hierfür sind divers und reichen von einer immer noch existierenden, mangelnden Akzeptanz bzw. Berührungsangst der Geisteswissenschaftler_innen hinsichtlich informationstechnischer Verfahren bis hin zu einer nicht an den Projektzielen ausgerichteten Implementierung der benötigten Software seitens der technischen Partner eines Projektes.

Ein bisher wenig berücksichtigter aber ganz zentraler Grund ist jedoch, dass die Ebene der Softwareentwicklung in den Nachhaltigkeitsdiskussionen der Digital

Humanities bisher kaum eine Rolle spielt. Erst seit diesem Jahr liegt ein erster Bericht zu generellen Voraussetzungen für die Nachhaltigkeit von Forschungssoftware vor (Hettrick 2016). Dieser kommt zu folgendem Schluss: „many researchers know how to code, but few understand the wider set of skills that are needed to develop reliable, reproducible and reusable software. […] software engineering should be incorporated […] at the very start of a research career.“ (Hettrick 2016, S. 14). Neben den sicherlich notwendigen Überlegungen zur Nachhaltigkeit geisteswissenschaftlicher Forschungsdaten sollte künftig mehr darauf geachtet werden, neben der reflexiven Ebene auch das konkrete entwicklerische Handwerkszeug in Digital Humanities Studiengänge einzubeziehen. Insbesondere müssen entwicklerische Leistungen als gleichrangige akademische Tätigkeit anerkannt werden (vgl. Hettrick, S. 13). Neben die Theorie sollte ein akademisch anerkanntes Digital Humanities “Craftsmanship” treten. [1]

Die Gründe für eine nicht nachhaltige Entwicklung geisteswissenschaftlicher Software sind relativ einfach zu identifizieren und keineswegs spezifisch für das akademische Entwicklungsumfeld. Sie spielen genauso in der freien Wirtschaft oder der Open Source Szene eine Rolle. Zu den Hauptgründen einer mangelnden Software-Nachhaltigkeit können beispielsweise gehören:

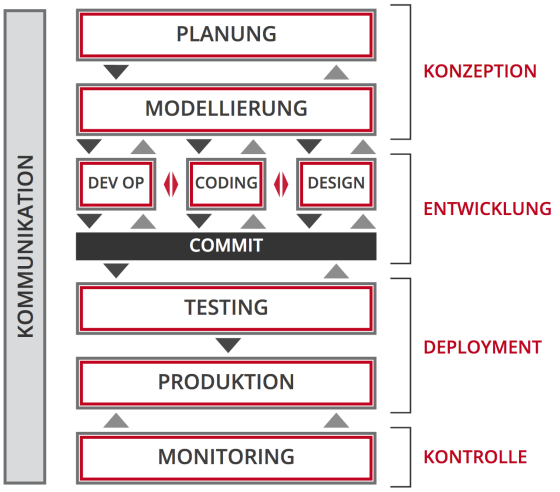
- Eine ausgelaufene Projektfinanzierung, wodurch der Weiterbetrieb der Software nicht mehr gewährleistet ist,
- Entwickler_innen, die dem Projekt nicht mehr zu Verfügung stehen, aber vor ihrem Weggang die ausschließlichen Wissensträger waren,
- eine veraltete und nicht mehr wartbare Infrastruktur,
- veralteter oder unverständlicher Programmcode, der von neu einsteigenden Entwickler_innen weitergeführt werden muss,
- Sicherheitslücken, die vermeidbar gewesen wären, jetzt aber einen Weiterbetrieb der Software verhindern,
- (schwerwiegende) Bugs in der Software, die erst im Produktivbetrieb auffallen, da vorher keine Softwaretests durchgeführt wurden,
- ein fehlendes Monitoring der geisteswissenschaftlichen Forschungsanwendung, wodurch Störfälle nicht oder erst spät auffallen.

Blickt man vor diesem Hintergrund in die freie Wirtschaft und Softwareindustrie und fragt nach aktuellen Projektmanagement-Methoden bzw. Herangehensweisen zur Steigerung der Qualität und Nachhaltigkeit einer Software, lässt sich sehr schnell feststellen, dass insbesondere die unter dem Stichwort „Agile Softwareentwicklung“ gefassten methodischen Ansätze sich sehr gut eignen, um den genannten Herausforderungen entgegenzutreten (vgl. Ayelt 2014). Obwohl agile Entwicklungsansätze häufig unterschiedliche Teilaspekte eines Entwicklungs-Workflows adressieren (bspw. die Konzeptionsebene, die Entwicklungsebene, die Ebene des Testings oder des Deployments einer Software), legen alle doch den Schwerpunkt auf eine kontinuierliche Kommunikation aller Projektbeteiligten untereinander (von den Stakeholdern über die Entwickler_innen bis zu den Testnutzer_innen und Endnutzer_innen). Agile Softwareentwicklung sieht häufig in einer für alle nachvollziehbaren Kommunikation den entscheidenden Schlüssel für ein erfolgreiches und nachhaltiges Softwareprodukt.

Hiermit befinden wir uns aber wiederum sehr nahe an den Digital Humanities. Schon lange wird für Digitale Geisteswissenschaftler_innen eine kommunikative Schlüsselstellung reklamiert. Als Mediatoren mit Fachwissen aus zwei Welten sollen sie eine für alle Parteien gemeinsame, verständliche Sprache entwickeln und so die unterschiedlichen geistes- und informationswissenschaftlichen Konzepte eines Forschungsprojektes miteinander in Einklang bringen.

Das Team der Digitalen Akademie der Mainzer Akademie der Wissenschaften und der Literatur integriert bereits seit 2009 Konzepte aus dem Bereich der agilen Softwareentwicklung in die tägliche Forschungs-, Entwicklungs- und Projektarbeit. Hierbei werden auf verschiedenen Ebenen Konzepte angewendet, die sich über die Zeit als besonders geeignet für geisteswissenschaftliche Anwendungskontexte herausgestellt haben. Sowohl zur Steigerung der Softwarequalität, insbesondere aber auch zur Steigerung der Nachhaltigkeit der Forschungsapplikationen wurde mit der Zeit eine an den Prinzipien der „Continuous Delivery“ ausgerichtete Prozesskette aufgebaut (zum Begriff vgl. Wolff 2015).

Die nachfolgende Grafik gibt einen Überblick über die einzelnen Ebenen dieser Prozesskette.



Auf Ebene der Konzeption und Programmierung kommen zwei Herangehensweisen zum Einsatz: das sogennannte Domain-Driven Design (DDD) und das Behaviour-Driven Development (BDD). Domain-Driven Design ist dabei zum einen eine Herangehensweise an die Modellierung komplexer Software, zum anderen ein bestimmtes Denkkonzept zur Steigerung der Produktivität von Softwareprojekten im Umfeld komplexer fachlicher Zusammenhänge. Das Hauptaugenmerk fällt dabei auf die Einführung einer ubiquitären (allgemein verständlichen) Sprache, welche in allen Bereichen der Softwareerstellung von den Konzeptionsgesprächen mit den Fachwissenschaftler_innen bis hin zur Code-Ebene verwendet werden sollte. Domain-Driven Design ist unabhängig von Programmiersprachen, Tools und Frameworks (vgl. Evans 2013, S. 13). DDD eignet sich ausgezeichnet für eine nachhaltige und offene Modellierung geisteswissenschaftlicher Anwendungskontexte, da iterativ gearbeitet wird. Zu Beginn der Domänen-Modellierung ist in geisteswissenschaftlichen Forschungsprojekten die eigentliche Datengrundlage und der Funktionsumfang der zu erstellenden Software häufig nicht vollständig klar. Beides entsteht sukzessive in der Beschäftigung mit dem Forschungsgegenstand. Somit können während der eigentlichen Entwicklung häufig neue Gegenstände auftauchen, noch nicht bedachte Eigenschaften hinzukommen oder sich auch Teile der Applikationslogik grundlegend ändern. Durch regelmäßige Iterationen nach dem DDD-Prinzip kann die Software kontinuierlich mit der sich stetig wandelnden Projektrealität verändern. Die Codebasis bleibt dabei im Einklang mit der Konzeptions- bzw. Modellierungsebene.

Behaviour-Driven Development wiederum geht davon aus, dass sich die Funktionalität einer Anwendungsdomäne (und somit die Geschäftslogik eines Domänen-Modells) durch formalisierte Szenarien in einer allgemeinverständlichen Sprache beschreiben lässt. BDD ist ein „outside-in“-Ansatz, der von außen (also mit dem Blick der Geisteswissenschaftlerinnen) auf eine Software blickt und deren Funktionalität in ausführbaren Tests dokumentiert. Ursprünglich im Umfeld des Test Driven Development entstanden, achtet auch BDD darauf, dass die Nutzungszenarien einer Software vor der eigentlichen Programmierung der Software erstellt werden. Dadurch dass die Tests in natürlicher Sprache nach einem festen Dreischritt-Prinzip (Angenommen…, Wenn…, Dann…) formuliert werden, kann die oftmals komplexe und wenig nachhaltige Präsentationsschicht und Funktionslogik einer Software in direkter Zusammenarbeit mit den Fachwissenschaftler_innen gemeinsam beschrieben und nachhaltig dokumentiert werden. In der Umsetzung hat dies für die Entwickler_innen den Vorteil, dass exakt nur soviel Code geschrieben werden muss, bis die jeweiligen Tests erfolgreich ablaufen und die Software exakt wie geplant funktioniert.

Als dritte wichtige Säule in einem nachhaltigen Entwicklungsprozess ist die Virtualisierung und Automation der Infrastruktur nach ‚DevOps‘-Prinzipien (eine Zusammenfügung der beiden Begriffe Development und Operations) zu nennen. ‚DevOps‘ betrachtet ‚Infrastruktur als Code‘ und setzt entsprechende Werkzeuge ein, um eine vollständige Kapselung der Softwareschicht und gleichzeitige Reproduzierbarkeit der Gesamtapplikation einschließlich ihrer Infrastruktur zu erreichen. Der große Vorteil dieser Verfahrensweise liegt in der automatisch enstehenden Dokumention hochgradig spezialisierter Anwendungsumgebungen. Gleichzeitiger sind die „Baupläne“ dieser Anwendungsumgebungen in einem Versionskontrollsystem versionierbar.

Alle bisher genannten Konzepte streben eine Versionierbarkeit ihrer Outputs an, was die Nachvollziehbarkeit und somit die Nachhaltigkeit auf der Software-Ebene deutlich steigert. Auf diese Weise hergestellte Software legt nicht nur offen, wie sie funktioniert, sondern wie sie hergestellt wurde und das auf allen Ebenen, von der Konzeption über die Programmierung und das Testing bis hin zum

Deployment. Insofern kommt dem ‚Commit', also dem wiederholten Einspielungsvorgang der jeweiligen Entwicklungsstände die Rolle des zentralen Dreh- und Angelpunktes einer nachhaltigen Softwareentwicklung zu. Zusammenfassend lässt sich im Abgleich zu den oben genannten Punkten festhalten, dass bei einer nachhaltigen Softwareentwicklung

- insbesondere die beständige Kommunikation aller Projektbeteiligten untereinander einen zentralen Faktor darstellt,
- ein gemeinsames Vokabular festgelegt werden und dies auf allen Ebenen konsequent angewendet werden muss (Konzeption, Datenschema, Code, Tests etc.).
- auf forschungsgetriebene, agile Entwicklungsmethoden gesetzt werden sollte,
- ein nachvollziehbarer Entwicklungsprozess durch Versionskontrolle gewährleistet werden muss,
- die Infrastruktur nach DevOps-Prinzipien virtualisiert und automatisiert werden sollte,
- Softwaretests vor jedem Live-Deployment durchzuführen sind,
- die Applikation im Produktivbetrieb kontinuierlich überwacht werden muss.

Innerhalb des Vortrags werden die dargelegten Konzepte und Methoden anhand von Projektbeispielen genauer illustriert und zur Diskussion gestellt. Der Beitrag versteht sich somit als ein Erfahrungsbericht aus der mehrjährigen Arbeit im Kontext der Digital Humanities Projekte der Digitalen Akademie der Mainzer Akademie sowie des Mainzer Zentrums für Digitalität in den Geistes und Kulturwissenschaften (mainzed).

Fußnoten

1. In Anlehnung an den Begriff des *software craftsmanship*.

Bibliographie

Komus, Ayelt (2014): *Status Quo Agile 2014*. Hochschule Koblenz. https://www.hs-koblenz.de/rmc/fachbereiche/wirtschaft/forschung-projekte-weiterbildung/forschungsprojekte/status-quo-agile/ [letzter Zugriff 26. August 2016].

Evans, Eric (2003): *Domain-Driven Design: Tackling Complexity in the Heart of Software*. Boston: Addison Wesley.

Hettrick, Simon (2016): *Research Software Sustainability: Report on a Knowledge Exchange Workshop*. Edinburgh: The Software Sustainability Institute.

Wolff, Eberhard (2015): *Continuous Delivery: Der pragmatische Einstieg*. Heidelberg: dpunkt.verlag.

Vernon, Vaughn (2013): *Implementing Domain Driven Design*. Boston: Addison Wesley.

Nachhaltigkeit als Prozess: Zur konzeptionellen Funktion digitaler Technologien in der Nachhaltigkeitssicherung für historische Fotos im Projekt efoto-Hamburg

Schumacher, Mareike
mareike.schumacher@uni-hamburg.de
Universität Hamburg, Deutschland

Abstract

efoto-Hamburg wird seit 2013 von der Universität Hamburg wissenschaftlich geleitet und von der Kulturbehörde der Stadt gefördert. Ziel ist der Aufbau einer gemeinsamen Bilddatenbank für private und behördliche Archive und Museen Hamburgs. Zugleich wird eine mobile App entwickelt, die die Bilddaten für die Öffentlichkeit zugänglich, nutz- und erfahrbar macht. Als ein zentrales Element verknüpfen Narrative Abgebildetes mit der Lebenswirklichkeit der Nutzer. Das Erzählen als Basis anthropologischer Überlieferung wird mit archivarischen Arbeitsweisen und informationstechnologischen Implementierungen verknüpft, um historisches Bildmaterial langfristig als Bestandteil einer lebendigen Stadtkultur zu erhalten. Unsere auf diesem Prinzip fußende interdisziplinär angelegte Nachhaltigkeitsstrategie möchte ich im hier vorgeschlagenen Vortrag erläutern und vor allen Dingen zur Diskussion stellen, welche Rolle digitale Technologien mit Blick auf Nachhaltigkeit als konzeptionelles kulturelles

Desiderat spielen können. Der Vortrag verknüpft die Diskussion um digitale Nachhaltigkeit mit Ansätzen aus der Kultur- und Erzähltheorie und zeigt ein Anwendungsbeispiel kultureller Nachhaltigkeit, welches die Verbindung von Wissenschaft und Öffentlichkeit anstrebt.

Nachhaltigkeit kultureller Daten: Konzeptionelles Desiderat und digitale Optionen

Unser grundlegendes Verständnis von Nachhaltigkeit basiert auf einer frühen Definition nachhaltiger Entwicklung aus dem Bericht der UN-Brundtland-Kommission “Our Common Future” (World Commission on Environment and Development 1987). Darin heißt es:“Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs.” (Ebd.: 41)

Nachhaltigkeit stellt sich hier vorrangig als Prozess der Vermittlung zwischen den Bedürfnissen der heutigen Gesellschaft und denen zukünftiger Generationen dar. Auch in Bezug auf das kulturelle Erbe wird dieses Spannungsfeld, das auch als Enkelgerechtigkeit (Die Bundesregierung 2015: 23) bezeichnet wird, als bedeutsam eingestuft (Willer 2013: 141).

In der frühen Definition der Nachhaltigkeit liegt der Schwerpunkt auf ökologischen Gesichtspunkten. Anschließend wurde das Konzept allerdings bereits um kulturelle Dimensionen erweitert. Für die efoto-Nachhaltigkeitsstrategie sind zwei Ansätze konzeptionell von besonderer Bedeutung:

Kulturelle Nachhaltigkeit: Kultur [1] wird neben Wirtschaft, Ökologie und Gesellschaft als Triebfeder für Nachhaltigkeit verstanden. Dabei geht es um die Frage, inwiefern Kultur behilflich sein kann, eine nachhaltige Entwicklung voran zu treiben. Nach diesem Verständnis kann Nachhaltigkeit nur gelingen, wenn diese in der Kultur eines sozialen Systems verankert ist. (Brocchi 2007)

Nachhaltigkeit der Kultur: Hier geht es hauptsächlich um die Frage, wie kulturelle Artefakte langfristig erhalten und lebendig gehalten werden können. Dabei spielen in unserem Kontext Strategien eine besondere Rolle, die das Nachhaltigkeitsdesiderat durch gezielte Nutzung digitaler Technologien zu beantworten suchen. Es eröffnen sich dabei unterschiedliche Problemfelder wie z.B.

Zugänglichkeit: Hier sind zwei Teilbereiche von überragender Bedeutung. Einerseits muss Zugang auf technischer Ebene geschaffen werden und erhalten bleiben. Objekte, die nicht ursprünglich digital sind, müssen digitalisiert und in Datenbanken und Portale überführt werden. In diesem Rahmen wird auch überlegt, wie digitale Langzeitarchivierung in Hard- und Software am besten zu leisten ist (Giebel 2013). Der zweite umfassend problematisierte Bereich ist der rechtliche Rahmen, der den öffentlichen Zugang meist erschwert (Steinhauer 2013). Die Vielfalt der Objektarten des kulturellen Erbes bedingt, dass häufig Rechte unterschiedlicher Art greifen.

Kuration: Grundsätzlich bietet die digitale Archivierung die Möglichkeit, Daten in großen Mengen zu erfassen. In Anbetracht der Masse der digitalen Artefakte stellt sich allerdings die Frage, was würdig ist, für das kulturelle Erbe bewahrt zu werden. Die bisher vorherrschende manuelle Sichtung durch Archivare und Kuratoren wird angesichts der schieren Datenmengen zunehmend unmöglich. Automatische digitale Kuration wurde als Lösung zwar formuliert (Zorich 2016: 14), birgt aber die Gefahr des menschlichen Kompetenzverlustes an den Computer bzw. an Algorithmen. Gleichzeitig scheint eine größere Scheu vor der digitalen Löschung als vor dem ‘Wegschmeißen’ von Artefakten zu bestehen, die als nicht archivwürdig klassifiziert werden. Das Löschen digitaler Daten wird oft als endgültig beschrieben, während ein weggeworfenes Objekt immer noch entweder physisch überdauern oder als private Kopie zu einem späteren Zeitpunkt wieder gefunden werden kann (Beinert; Straube 2013: 28f).

Authentizität: Dieser Punkt bezieht sich hauptsächlich auf ursprünglich digitale Artefakte, wie Zeugnisse des digitalen Wandels selbst (z. B. Webseiten). Aufgrund der Verankerung in Hard- und Software, die oft schnell obsolet wird, sind die sogenannten “born digital” Artefakte oftmals schon frühzeitig nicht mehr auf die gleiche Weise aufrufbar wie zu der Zeit, als sie entwickelt wurden (Crueger 2013). Aber auch in Bezug

auf grafische Darstellungen von Digitalisaten stellt sich die Frage, inwiefern technisch langfristig gewährleistet werden kann, dass kulturelle Artefakte zumindest ähnlich betrachtet werden können wie ihre analogen Pendants (Fröhlich 2013).

Eine Nachhaltigkeitsstrategie für efoto-Hamburg

In den Partnerinstitutionen [2] von efoto-Hamburg liegen insgesamt mehrere Millionen Bilddaten in analoger und digitaler Form vor, die unterschiedlich gut zugänglich sind. In einem ersten Schritt wird der Import einer Teilmenge von rund 100.000 Datensätzen aus fünf Partnerinstitutionen angestrebt, die in der mobilen App auf einer interaktiven Karte zugänglich gemacht werden. Darüber hinaus ist es eines der Projektziele, nach und nach möglichst viele der digital vorliegenden Bilder so bereitzustellen, dass sie in die digitale Nachhaltigkeit überführt werden können. Die Nachhaltigkeitsstrategie von efoto-Hamburg umfasst zwei einander ergänzende Vorgehensweisen.

Strategische Dimension: Diskursives Kulturkonzept und narrative Struktur

Kulturelle Nachhaltigkeit muss nach unserem Verständnis als Entwicklungsprozess aufgefasst werden, der kulturelle Artefakte nicht nur im Sinne des kulturellen Erbes an Folgegenerationen übergibt, sondern von Beginn an eine Art *gelebte* Enkelgerechtigkeit unterstützt. Erst Kultur als tatsächliche Reflexionspraxis macht die kulturellen Artefakte auch diskursiv funktional; sie etabliert somit eine Diskursstruktur, innerhalb derer prinzipiell jeder als Zeitzeuge agieren und seine Eindrücke festhalten und teilen kann. In diesem Kontext können kulturelle Artefakte zugleich Anlass wie Gegenstand diskursiver und reflexiver Prozesse werden.

Diese strategische Prämisse von efoto-Hamburg ruht auf zwei konzeptionellen Grundpfeilern. Der erste ist ein an Luhmann angelehntes Kulturverständnis: Kultur wird als ein Prozess verstanden, der sich auf drei Ebenen abspielt; der Objektebene, der Reflektion erster und der Reflektion zweiter Ordnung. Erst wenn alle drei Ebenen miteinander verknüpft sind, ist die Voraussetzung dafür gegeben, dass ein Artefakt nachhaltig im Kulturprozess verankert sein kann. (Luhmann 2011: 140 und Luhmann 1999: 99) Eine Verknüpfung der drei Ebenen könnte z.B. wie folgt aussehen:

Die zweite Säule ist die narrative Natur dieses Kulturprozesses. Grundannahme ist hier, dass auch in der digitalisierten Gesellschaft Überlieferung nur durch narrative Kommunikation gewährleistet werden kann. Identitäten von Individuen und Gruppen werden durch Minimalnarrative, sogenannte Small Stories (Bamberg; Georgakopoulou 2008 und Georgakopoulou 2007), ausgeformt. Die Motivation des oben abgebildeten Nutzers A ist demnach identitätsbildender Natur. Er oder sie verknüpft das, was im kulturellen Artefakt dargestellt ist, mit einem Ereignis oder einem Teilaspekt aus der eigenen Lebensgeschichte oder seiner Persönlichkeit, um diese innerhalb der efoto-Community zu stärken und/oder auf ähnliche Persönlichkeiten und/oder Lebensgeschichten zu treffen. Nutzer B und C sind ähnlich motiviert, auch wenn ihre Kommunikation nicht durch das Objekt, sondern durch die Reflexion über dasselbe ausgelöst wird. Indem alle drei Beispielnutzer über Small Stories die historischen Bilddaten mit ihrer Lebenswirklichkeit verknüpfen, halten sie diese lebendig. Artefakte, die so in einem aktuellen Kulturprozess verankert werden, sind für efoto besonders bewahrenswert. Diese Bewertung erfolgt also dynamisch in einem sozio-kulturellen System. Damit ergibt sich nun die Frage, welche spezifische Rolle digitalen Technologien in diesem Zusammenhang zukommt.

Die Rolle digitaler Technologien im Kontext der Nachhaltigkeitsstrategie

Im Vordergrund von eFoto-Hamburg steht weder ein archivarisches Interesse noch eine Kulturvermittlung als Überzeugungsarbeit: nicht das möglichst nachhaltige ‘Aufbewahren’ historischer Bilder im digitalen Format und auch nicht das Erzeugen kultureller Akzeptanz für diese Bilder ist die raison d’etre des Projekts, sondern das Einbinden der Bilder in aktuelle reflexive Prozesse.

Die manuelle Einzelprüfung durch Kuratoren, die in den Partnerinstitutionen von efoto-Hamburg bereits stattgefunden hat, bevor die Bilddaten auf die digitale Plattform gelangen,

versteht sich daher als Vorstufe, die die Interaktion der Community mit und über die digitalisierten Artefakte vorbereitet und unterstützt. Jeder Nutzer gilt unabhängig von Faktoren wie z.B. seinem Alter als Zeitzeuge und wird als solcher in den weiteren Kurationsprozess einbezogen. Auf diese Weise nehmen Nutzer generationenübergreifend an der Entwicklung dieses Teilbestandes des kulturellen Erbes teil - und an eben dieser Stelle bieten digitale Technologien nun die Möglichkeit, die Nachhaltigkeit kultureller Artefakte nicht nur im Sinne eines statischen (archivarischen) 'Vorhaltens' digitaler Repräsentationen zu sichern, sondern vielmehr auch im Sinne eines 'Lebendighaltens' durch den aktiven Gebrauch und die Einbettung in gelebte kulturelle Diskursprozesse zu befördern.

Im Projekt efoto wird zu diesem Zweck ein System aus konkreten technischen Features entwickelt, welches Nutzern unterschiedliche Möglichkeiten eröffnet, die im digitalen Format vorliegenden Bilder in aktive Gebrauchsprozesse einzubinden und so *kulturelle Nachhaltigkeit qua kultureller Nutzung* zu sichern. Die mobile efoto-App umfasst Features wie Stadtrundgänge, Zeitzeugen-Interviews, ein Kommentar-System, einen Bildrechte-Wegweiser, eine interaktive Karte oder das "historische Selfie". Anhand dieser Beispielfeatures wird im vorgeschlagenen Vortrag erläutert werden, welche konzeptionellen Ideen in die Entwicklung eingeflossen sind und wie diese die Nachhaltigkeitsstrategie umsetzen.

Relevanz und Anschlussfähigkeit

Der vorgeschlagene Beitrag versteht sich als Anwendungsbeispiel einer interdisziplinär ausgerichteten kulturellen Nachhaltigkeitsstrategie. Ansätze aus der Erzähltheorie, der Kulturwissenschaft und den Museumswissenschaften werden mit einem ökologisch-politischen Verständnis von Nachhaltigkeit verbunden. Damit versucht efoto-Hamburg zu erproben, was bisher sowohl im wissenschaftlichen als auch im gesellschaftlichen Diskurs lediglich theoretisch reflektiert und teilweise auch proklamiert worden ist: eine kulturell angetriebene Nachhaltigkeit kulturellen Datenmaterials, deren Kernidee die Einbindung von Artefakten in (digital unterstützte) Gebrauchs- und Reflexionsprozesse und nicht deren bloßes langfristiges 'Bewahren' in möglichst stabilen medialen Formaten ist.

efoto-Hamburg ist dabei nicht nur interdisziplinär ausgerichtet, sondern bezieht auch Bürger und Besucher der Stadt auf allen oben skizzierten Ebenen des Kulturprozesses ein. Damit verbindet das Projekt Wissenschaft und Öffentlichkeit in möglichst durchlässiger Weise. Für die digitalen Geisteswissenschaften stellt das Projekt nicht nur ein Anwendungsbeispiel dar, sondern eine lebendige Plattform, die für zahlreiche Anschlussuntersuchungen offen ist.

Fußnoten

1. Kultur wird hier - abweichend vom im Folgenden erläuterten Kulturverständnis innerhalb des Projektes efoto-Hamburg - sehr umfassend als Wechselspiel der Einwirkung des Menschen auf seine Umwelt und der Einwirkung der Umwelt auf den Menschen verstanden.
2. Dazu gehören das Hamburger Staatsarchiv, das Landesamt für Geoinformation und Vermessung, die Hamburger Geschichtswerkstätten, das Museum der Arbeit, das Museum für Kunst und Gewerbe, das Polizeimuseum und die Hamburger Feuerwehrhistoriker.

Bibliographie

Bamberg, Michael / Georgakopoulou, Alexandra (2008): „Small stories as a new perspective in narrative and identity analysis", in: De Fina, Anna / Georgakopoulou, Alexandra (eds.): *Narrative Analysis in the Shift from Texts toPractices*. Special Issue of *Text & Talk 28*: 377–396.

Beinert, Tobias / Straube, Armin (2013): „Aktuelle Herausforderungen der digitalen Langzeitarchivierung", in: Klimpel, Paul / Keiper, Jürgen (eds.): *Was bleibt? Nachhaltigkeit der Kultur in der digitalen Welt*. Berlin: iRights Media Verlag 27–46.

Brocchi, Davide (2007): „Die kulturelle Dimension der Nachhaltigkeit", in: *Magazin Cultura21* http://davidebrocchi.eu/wp-content/uploads/2013/08/2007_dimension_nachhaltigkeit.pdf [letzter Zugriff 25. August 2016].

Bundesregierung (2015): *Meilensteine der Nachhaltigkeitspolitik. Weiterentwicklung der nationalen Nachhaltigkeitsstrategie* http://www.bundesregierung.de/Content/DE/_Anlagen/2015/02/2015-02-03-meilensteine-der-nachhaltigkeitspolitik.pdf?

__blob=publicationFile [letzter Zugriff 17. August 2016].

Crueger, Jens 2013): „Die Dark Ages des Internet“, in: Klimpel, Paul / Keiper, Jürgen (eds.): *Was bleibt? Nachhaltigkeit der Kultur in der digitalen Welt*. Berlin: iRights Media Verlag 191–198.

Fröhlich, Jan (2013): „Farbraum und Bildzustand im Kontext der Langzeitarchivierung“, in: Klimpel, Paul / Keiper, Jürgen (eds.): *Was bleibt? Nachhaltigkeit der Kultur in der digitalen Welt*. Berlin: iRights Media Verlag 119–125.

Georgakopoulou, Alexandra (2007): „Small Stories, Interaction and Identities“, in: *Studies in Narrative 8*. Amsterdam / Philadelphia: John Benjamins.

Giebel, Ralph (2013): „Speichertechnologie und Nachhaltigkeit“, in: Klimpel, Paul / Keiper, Jürgen (eds.): *Was bleibt? Nachhaltigkeit der Kultur in der digitalen Welt*. Berlin: iRights Media Verlag 95–108.

Luhmann, Niklas (1999): *Die Kunst der Gesellschaft*. Frankfurt am Main: Suhrkamp.

Luhmann, Niklas (2011): *Einführung in die Systemtheorie*. Heidelberg: Carl Auer.

Steinhauer, Eric (2013): „Wissen ohne Zukunft? Der Rechtsrahmen der digitalen Langzeitarchivierung von Netzpublikationen“, in: Klimpel, Paul / Keiper, Jürgen (eds.): *Was bleibt? Nachhaltigkeit der Kultur in der digitalen Welt*. Berlin: iRights Media Verlag 61–80.

Willer, Stefan (2013): „Kulturelles Erbe und Nachhaltigkeit“, in: Klimpel, Paul / Keiper, Jürgen (eds.): *Was bleibt? Nachhaltigkeit der Kultur in der digitalen Welt*. Berlin: iRights Media Verlag 139–152.

World Commission on Environment and Development (1987): *Our Common Future*. http://www.un-documents.net/our-common-future.pdf [letzter Zugriff 17. August 2016].

Zorich, Diane (2015): *Report of the Summit in Digital Curation in Art Museums* http://advanced.jhu.edu/wp-content/uploads/2016/04/digitalCuration_summitReport10_2015.pdf [letzer Zugriff 24. August 2016].

Netzwerkdynamik, Plotanalyse – Zur Visualisierung und Berechnung der ›progressiven Strukturierung‹ literarischer Texte

Trilcke, Peer
trilcke@uni-potsdam.de
Universität Potsdam, Deutschland

Fischer, Frank
ffischer@hse.ru
Higher School of Economics, Moskau, Russland

Göbel, Mathias
goebel@sub.uni-goettingen.de
Staats- und Universitätsbibliothek Göttingen, Deutschland

Kampkaspar, Dario
kampkaspar@hab.de
Herzog-August-Bibliothek Wolfenbüttel, Deutschland

Kittel, Christopher
contact@christopherkittel.eu
Universität Graz, Österreich

Forschungsstand

Die Anwendung von Methoden der Netzwerkanalyse auf literarische Texte hat sich in den letzten Jahren zu einem eigenständigen Forschungsfeld der *Digital Literary Studies* entwickelt. Im Vordergrund stehen dabei häufig computerlinguistische Fragen, insbesondere solche nach der automatisierten Extraktion von Netzwerkdaten (z.B. qua *named entity recognition*, *co-reference resolution*) und deren Evaluation (u. a. Elson et al. 2010; Park et al. 2013; Agrarwal et al. 2013; Rochat 2014; Fischer et al. 2015; Waumans et al. 2015; Jannidis et al. 2016).

Darüber hinaus wird ausgelotet, inwiefern sich mittels visueller und/oder

statistischer Auswertung der Netzwerkdaten genuin literaturwissenschaftliche Erkenntnisse gewinnen bzw. neue Wege der literaturwissenschaftlichen Analyse entwickeln lassen: Neben Ansätzen zur quantitativen Beschreibung und Hierarchisierung des Figurenpersonals (Jannidis et al. 2016) werden hier, im Rahmen korpusbasierter Analysen, Optionen der literaturhistorischen Periodisierung auf Basis von quantitativen Strukturdaten diskutiert (Trilcke et al. 2015) sowie Typen der ästhetischen Modellierung sozialer Formationen in und durch literarische Texte differenziert (Stiller et al. 2003; Stiller & Hudson 2005; Trilcke et al. 2016).

Forschungsdesiderat: Plotanalyse

Nahezu keine Rolle spielte bisher jedoch ein durchaus hehres Erkenntnisversprechen, das – bereits in der prä-automatisierten Zeit formuliert (de Nooy 2006) – auch den Fluchtpunkt des einschlägigen ›Pamphlets‹ von Franco Moretti steht: dass nämlich die Netzwerkanalyse als ein Instrumentarium der quantitativen »plot analysis« (Moretti 2011) fungieren könne.

Tatsächlich lässt sich dieses Erkenntnisversprechen mit den derzeit verfolgten Ansätzen im Bereich der literaturwissenschaftlichen Netzwerkanalyse kaum aufgreifen, geschweige denn einlösen (so auch Prado et al. 2016). Denn die sequentielle Dimension literarischer Texte, mithin ihre Temporalität, bleibt hier in der Regel ausgeblendet: Erfasst, visualisiert und analysiert werden statische Netzwerke. Plot ist jedoch wesentlich ein Konzept, das die Temporalität narrativer (wie auch dramatischer) [1] Texte theoretisch fassen soll: »the repeated attempts to redefine parameters of plot reflect both the centrality and the complexity of the temporal dimension of narrative« (Dannenberg 2005: 435). Plot lässt sich begreifen als Konzept zur Beschreibung der »progressive structuration« (Kukkonen 2013, §4) literarischer Texte.

Versuche, die Netzwerkanalyse in Richtung einer quantitativen Plotanalyse weiterzuentwickeln, stehen also zunächst vor der Aufgabe, bei ihrer Modellierung des Untersuchungsgegenstandes die Zeitdimension zu berücksichtigen. Der Text ist entsprechend nicht lediglich als ein statisches Netzwerk zu modellieren, sondern als eine sich über die Zeit verändernde Folge von Netzwerkzuständen. Erst anhand dieser Netzwerkdynamiken lassen sich die Erkenntnispotenziale, die netzwerkanalytische Zugänge für die quantitative Plotanalyse bergen, überhaupt diskutieren.

Forschungsvorhaben

Der projektierte Vortrag wird – in Anschluss an Prado et al. 2016 – aus theoretischer und methodischer Perspektive sowie anhand exemplarischer Fallstudien eine Erweiterung der bisherigen, auf die Analyse *statischer Strukturen* fokussierten Forschung zu literarischen Netwerken um die Analyse *progressiver Strukturierungen* vorschlagen. Übergreifendes Ziel ist es, zu prüfen, ob (und mit welchen Einschränkung) sich auf diesem Wege ein Beitrag zur Operationalisierung des literaturwissenschaftlichen Plot-Konzepts erarbeiten lässt. Dabei soll es nicht darum gehen, das semantische reiche und vielseitige Plot-Konzept der ›traditionellen‹ Literaturwissenschaft durch ein quantitatives und insofern notgedrungen reduktionistisches Konzept zu ersetzen. Vielmehr soll zunächst der wesentlich bescheidenere Nachweis erbracht werden, dass sich bestimmte Aspekte dessen, was gemeinhin im Rahmen des Plot-Konzepts diskutiert wird, durchaus mittels der computerbasierten Analyse von Netzwerkdynamik beobachten lassen, etwa ereignishafte Konfliktverläufe (so schon Moretti 2011), Formen der sozialen Integration und Desintegration von Figuren oder basale Techniken der Handlungsführung, z.B. die Komposition von Haupt- und Nebenhandlung(en).

Entsprechend der zweigleisigen Auswertungsroutinen, die auf netzwerkanalytische Daten angewendet werden, wird der Vortrag zwei Szenarien der netzwerkbasierten Analyse der progressiven Strukturierung literarischer Texte diskutieren: zum einen (3.1) sind Möglichkeiten und Erkenntnispotenziale der *Visualisierung* dynamischer Netzwerke, zum anderen (3.2) Möglichkeiten und Erkenntnispotenziale der Berechnung *netzwerkanalytischer Maße* für dynamische Netzwerke auszuloten.

Visualisierung von Netzwerkgraphen

Während die Visualisierung dynamischer Netzwerke in anderen Domänen bereits seit längerem gang und gäbe ist (vgl. exemplarisch Pohl et al. 2008; Frederico et al. 2011), wurde erst vor Kurzem der Versuch unternommen, entsprechende Visualisierungsverfahren auch auf literarische Netzwerke anzuwenden (Xanthos et al. 2016). Während Xanthos et al. u.a. auf didaktische Anwendungsszenarien hinweisen, wird ein literaturwissenschaftliches Erkenntnispotenzial lediglich angedeutet; eine Diskussion dessen, was durch eine solche Visualisierung nicht nur *sichtbar*, sondern auch *erkennbar* wird, bleibt aus.

Hingegen zeigen erste, im Vortrag zu vertiefende Zwischenergebnisse unserer Analysen, dass die dynamische Visualisierung insbesondere dann erkenntnisrelevant wird, wenn es darum geht, multiplexe Netzwerke zu modellieren, d. h. Netzwerke, die unterschiedliche Interaktionstypen zugleich erfassen. So zeigt eine statische Visualisierung von Lessings bürgerlichem Trauerspiel *Emilia Galotti* die Familie Galotti als eine geschlossene Triade (siehe Abb. 1): Die Kanten symbolisieren hier szenische Kopräsenzen (Interaktionstyp 1), wobei jene Kanten, die *zugleich* Verwandtschaftsverhältnisse darstellen, rot erscheinen (Interaktionstyp 2).

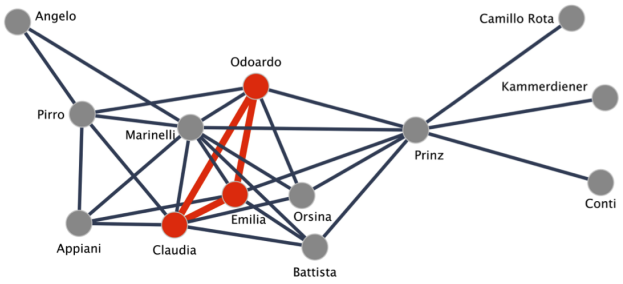


Abb. 1: Statisches Netzwerk zu Lessing: *Emilia Galotti* (rote Knoten: Familienmitglieder; rote Kanten: Familienmitglieder sind szenisch kopräsent)

Zerlegt man das statische Dramennetzwerk (Abb. 1) nun nach Akten und dynamisiert es damit, so zeigt sich, dass die Familie Galotti zu keinem Zeitpunkt des Dramas gemeinsam auf der Bühne steht (vgl. Abb. 2).

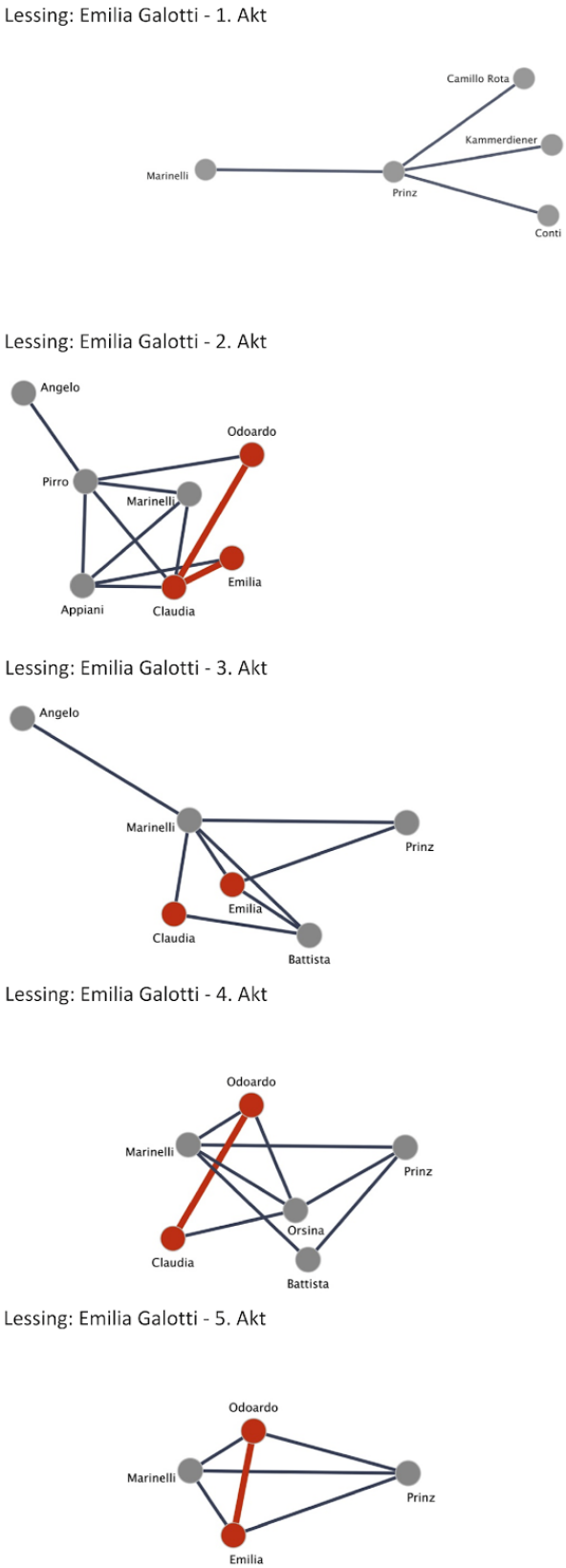


Abb. 2: Dynamisches Netzwerk zu Lessings *Emilia Galotti,* zerlegt nach Akten

Anschaulich und *erkennbar* wird auf diese Weise eine Position der traditionellen Forschung, nach der Lessing in *Emilia Galotti* nicht nur die äußere Bedrohung der ›bürgerlichen‹ Kleinfamilie, sondern auch deren innere Problematik inszeniert hat (siehe z. B. Alt 1994: 268). Die Analyse der dynamischen Strukturierung zeigt hier die soziale Desintegration der familiäre Triade, die als formal beschreibbarer Teilaspekt des zentralen dramatischen Konflikts verstanden werden kann.

Dass dynamische Visualisierungen in diesem Sinne aus literaturwissenschaftlicher Sicht v.a. für die Analyse multiplexer Netzwerke produktiv gemacht werden können, werden wir im Vortrag anhand weiterer Beispiele aus dem dlina-Korpus (philologisch kuratierte Netzwerkdaten zu 465 deutschsprachige Dramen aus der Zeit 1730–1930, siehe https://dlina.github.io/Introducing-DLINA-Corpus-15-07-Codename-Sydney/) zeigen. Darüber hinaus werden wir zum Zweck eines intergenerischen Vergleichs exemplarisch dynamische Visualisierung von Romannetzwerken diskutieren. Zu reflektieren sind hier insbesondere Fragen der Sequenzierung: Während Dramen mit ihrer Einteilung in Akte und Szenen eine naheliegende Segmentierung vorgeben, liefert die romantypische Einteilung in Kapitel keine vergleichbar überzeugenden Ergebnisse.

Berechnung netzwerkanalytische Maße

Mehr noch als die Visualisierung statischer Netzwerke stellt diejenige dynamischer im Grunde keine Option eines korpusbasierten *distant reading* dar. Sie ermöglicht zwar die anschauliche Modellierung einzelner Netzwerke, kann aber nur begrenzt Erkenntnisse über eine große Anzahl von Netzwerken liefern: Methoden, mit denen sich die auf algorithmischen Layouts basierenden Netzwerkgraphen kontrolliert miteinander vergleichen lassen, fehlen weitgehend; zudem kostet die Rezeption von dynamischen Visualisierungen – etwa der von Xanthos et al. 2016 präsentierten Prototypen – schlicht Zeit, wir haben es hier also eher mit *fast reading,* denn mit *distant reading* zu tun. Die Berechnung netzwerkanalytischer Maße und deren statistische Weiterverarbeitung bietet hingegen Möglichkeiten, aus einer dezidierten *distant reading*-Perspektive sowohl allgemeine Charakteristika der Netzwerke eines Korpus zu beschreiben als auch, vergleichend, spezifische formale Typen von Netzwerken innerhalb des Korpus zu identifizieren (entsprechend unserer Überlegungen zum *Small World*-Phänomen in statischen Netzwerken, siehe Trilcke et al. 2016).

Von Carley (2003: 135–136) wurden dabei mehrere rudimentäre globale Maße (i. e. *size*, *density*, *homogeneity in the distribution of ties*, *rate of changes in nodes*, *rate of changes in ties*) für die Analyse dynamischer Netzwerke vorgeschlagen. Darüber hinaus haben Prado et al. 2016 für die Anwendung von akteursorientierten Maßen, v.a. Zentralitätsindices, bei der Rekonstruktion von Plot -Verläufen plädiert. Im Vortrag werden wir einzelne dieser Maße – u. a. *size* pro Akte und Szenen; *density* pro Akte und Szene; die *change-rates*; sowie einfache Zentralitätsmaße – für das dlina-Korpus berechnen; die dafür nötigen Daten liegen bereits, philologisch kuratiert, in den dlina-Zwischenformat-Dateien vor (zum Zwischenformat: https://dlina.github.io/Introducing-Our-Zwischenformat/ – die Daten sind offen, siehe unser Github-Repositorium: https://github.com/dlina); eine entsprechende Erweiterung des in Python geschriebenen Auswertungstools *dramavis* (Kittel / Fischer 2016) wird derzeit entwickelt. Die erhobenen Daten werden wir schließlich mit Rekurs auf ausgewählte literaturwissenschaftliche Konzepte für die Beschreibung spezifischer Plot-Phänomene diskutieren, insbesondere in Hinblick auf Expositionstypen (Pfister 1977: 124–136), auf die ›klassische‹ Aktstruktur der Tragödie sowie auf das Kompositionsprinzip von Haupt- und Nebenhandlung (Pfister 1977: 286–289).

Resümee

Der Vortrag liefert einen Beitrag zur Methodenentwicklung und -reflektion im Bereich der *Digital Literary Studies.* Auf literaturtheoretisch-methodologischer Ebene diskutiert er Möglichkeiten einer netzwerkanalytischen Operationalisierung des literaturwissenschaftlichen Plot -Konzepts, wobei der literarische Text zu diesem Zweck nicht, wie bisher die Regel, als statische Struktur, sondern als ›progressive Strukturierung‹ modelliert wird. Als empirische Grundlage der Methodendiskussion fungieren Analysen von Dramen und Romanen, in denen exemplarisch die Potenziale und die Grenzen des Ansatzes verdeutlich werden.

Fußnoten

1. Unter systematischen Gesichtspunkten können die Unterschiede zwischen narrativen und dramatischen Texten in Hinblick auf das Plot-Konzept zunächst vernachlässigt werden (vgl. Korthals 2003); entsprechend wurden sowohl ›epische‹ als auch ›dramatische‹ Texte bis ins 19. Jahrhundert hinein verschiedentlich unter dem Oberbegriff ›pragmatische Gattung‹ vereint.

Bibliographie

Agarwal, Apoorv / Corvalan, Augusto / Jensen, Jacob / Rambow, Owen (2012): „Social Network Analysis of Alice in Wonderland“, in: *Proceedings of the Workshop on Computational Linguistics for Literature*. Montréal 88–96 http://www.aclweb.org/anthology/W12-2513 [letzter Zugriff 25. August 2016].

Alt, Peter-André (1994): *Die Tragödie der Aufklärung*. Eine Einführung. Tübingen / Basel: Francke.

Carley, Kathleen M. (2003): „Dynamic Network Analysis“, in: Breiger, Ronald / Carley, Kathleen M. / Pattison, Philipp (eds.): *Dynamic Social Network Modeling and Analysis*. Workshop Summary and Papers. Washington D.C.: 133–145 http://www.nap.edu/read/10735/chapter/9 .

Dannenberg, Hilary (2005): „Plot“, in: Herman, David / Jahn, Manfred / Ryan, Marie-Laure (eds.): *The Routledge Encyclopedia of Narrative Theory*. London: Routledge 435–439.

Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario / Trilcke, Peer (2015): „Digital Network Analysis of Dramatic Texts“, in: *DH2015: Global Digital Humanities* http://dh2015.org/abstracts/xml/FISCHER_Frank_Digital_Network_Analysis_of_Dramati/FISCHER_Frank_Digital_Network_Analysis_of_Dramatic_Text.html [letzter Zugriff 25. August 2016].

de Nooy, Wouter (2006): „Stories, Scripts, Roles, and Networks“, in: *Structure and Dynamics* 1.2 http://escholarship.org/uc/item/8508h946#page-1 [letzter Zugriff 25. August 2016].

Elson, David K. / Dames, Nicholas / McKeown, Kathleen R. (2010): „Extracting Social Networks from Literary Fiction“, in: *Proceedings of ACL-2010*. Uppsala: 138–147 http://dl.acm.org/ft_gateway.cfm?id=1858696&type=pdf&CFID=659731302&CFTOKEN=83466756 [letzter Zugriff 25. August 2016].

Federico, Paolo / Aigner, Wolfgang / Miksch, Silvia / Windhager, Florian / Zenk, Lukas (2011): „A Visual Analytics Approach to Dynamic Social Networks“, in: *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW)*. Graz http://publik.tuwien.ac.at/files/PubDat_198995.pdf [letzter Zugriff 25. August 2016].

Jannidis, Fotis / Reger, Isabella / Krug, Markus / Weimer, Lukas / Macharowsky, Luisa / Puppe, Frank (2016): „Comparison of Methods for the Identification of Main Characters in German Novels“, in: *DH2016: Conference Abstracts* 578–582 http://dh2016.adho.org/abstracts/297 [letzter Zugriff 25. August 2016].

Kittel, Christopher / Fischer, Frank (2016): *dramavis (v0.2.1)*. GitHub https://github.com/lehkost/dramavis [letzter Zugriff 25. August 2016].

Korthals, Holger (2003): *Zwischen Drama und Erzählung*. Ein Beitrag zur Theorie geschehensdarstellender Literatur. Berlin: Erich Schmidt

Kukkonen, Karin (2013): „Plot“, in: Hühn, Peter et al. (eds.): *The Living Handbook of Narratology*. Hamburg http://www.lhn.uni-hamburg.de/article/plot [letzter Zugriff 25. August 2016].

Moretti, Franco (2011): *Network Theory, Plot Analysis* (= Stanford Literary Lab Pamphlets, No. 2). 1.5.2011. http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf [letzter Zugriff 25. August 2016].

Park, Gyeong-Mi / Kim, Sung-Hwan / Cho, Hwan-Gue (2013): „Structural Analysis on Social Network Constructed from Characters in Literature Texts“, in: *Journal of Computers* 8.9: 2442–2447 http://ojs.academypublisher.com/index.php/jcp/article/view/jcp080924422447/7672 [letzter Zugriff 25. August 2016].

Pfister, Manfred (1977): *Das Drama*. Theorie und Analyse. München: Fink.

Pohl, Mathias / Reitz, Florian / Birke, Peter (2008): „As Time Goes by. Integrated Visualization and Analysis of Dynamic Networks“, in: *AVI 2008 – Proceedings of the Working Conference on Advanced Visual Interfaces*. Neapel 372–375 http://doi.acm.org/10.1145/1385569.1385636 [letzter Zugriff 25. August 2016].

Prado, Sandra D. / Dahmen, Silvio R. / Bazzan, Ana L.C. / Carron, Padraig Mac / Kenna, Ralph (2016): „Temporal Network Analysis of Literary Texts“, 24.2.2016 https://arxiv.org/pdf/1602.07275 [letzter Zugriff 25. August 2016].

Rochat, Yannick (2014): *Character Networks and Centrality*. Thèse de Doctorat. Lausanne https://infoscience.epfl.ch/record/203889/files/yrochat_thesis_infoscience.pdf [letzter Zugriff 25. August 2016].

Stiller, Jaames / Nettle, Daniel / Dunbar, Robin I. M. (2003): „The Small World of Shakespeare's Plays“, in: *Human Nature* 14: 397–408 https://www.staff.ncl.ac.uk/daniel.nettle/shakespeare.pdf [letzter Zugriff 25. August 2016].

Stiller, James / Hudson, Mathew (2005): „Weak Links and Scene Cliques Within the Small World of Shakespeare“, in: *Journal of Cultural and Evolutionary Psychology* 3: 57–73.

Trilcke, Peer / Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario (2015): „200 Years of Literary Network Data“ [Blogposts], https://dlina.github.io/200-Years-of-Literary-Network-Data/ [letzter Zugriff 25. August 2016].

Trilcke, Peer / Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario / Kittel, Christopher (2016): „Theatre Plays as ›Small Worlds‹? Network Data on the History and Typology of German Drama, 1730-1930“, in: *DH2016: Conference Abstracts* 417–419 http://dh2016.adho.org/abstracts/407 [letzter Zugriff 25. August 2016].

Waumans, Michaël C. / Nicodème, Thibaut / Bersini, Hugues (2015): „Topology Analysis of Social Networks Extracted from Literature“, in: *Plos One* 3. Juni 2015 10.1371/journal.pone.0126470.

Xanthos, Aris / Pante, Isaac / Rochat, Yannick / Grandjean, Martin (2016): „Visualising the Dynamics of Character Networks“, in: *DH2016: Conference Abstracts* 417–419 http://dh2016.adho.org/abstracts/407 [letzter Zugriff 25. August 2016].

Niklas Luhmanns Werk- und Lesekosmos - DH in der bibliographischen Dimension

Goedel, Martina
mgoedel@uni-koeln.de
Cologne Center for eHumanities, Universität zu Köln, Deutschland

Zimmer, Sebastian
sebastian.zimmer@uni-koeln.de
Cologne Center for eHumanities, Universität zu Köln, Deutschland

Der hier vorgestellte Workflow für die Digitalisierung und Integration bibliographischer Informationen ist Teil des Forschungsprojektes Niklas Luhmann - Theorie als Passion. Wissenschaftliche Erschließung und Edition des Nachlasses . Das Langzeitvorhaben (2015-2030) an der Fakultät für Soziologie der Universität Bielefeld in Kooperation mit dem Cologne Center for eHumanities (CCeH) wird im Akademienprogramm durch die Nordrhein-Westfälische Akademie der Wissenschaft und der Künste gefördert. Weitere Kooperationspartner sind das Archiv und die Bibliothek der Universität Bielefeld. [1]

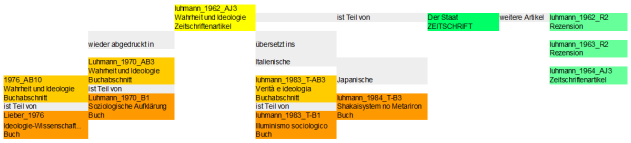
Ziel des Gesamtprojektes ist die Sicherung, Digitalisierung, Erschließung, werkgenetische Erforschung und teilweise Edition des wissenschaftlichen Nachlasses Niklas Luhmanns. Zu diesem Zweck werden die bewahrenswerten Teile des Nachlasses (Manuskripte, Zettelkasten, Korrespondenz, Bibliothek etc.) zunächst archivarisch gesichert und in den Teilen, die wissenschaftlich erschlossen werden sollen, digitalisiert, sowie für die weitere Bearbeitung bereitgestellt. Die daran anschließende Edition will den Luhmannschen Nachlass als geistesgeschichtliches Dokument der wissenschaftlichen Forschung sowie der interessierten Öffentlichkeit zugänglich machen. Sie bildet dadurch zugleich die Grundlage für die Entwicklung einer kritisch gesicherten infrastrukturellen Wissensressource, auf welche die interdisziplinäre und internationale Forschung zur und mit der Theorie Luhmanns zukünftig zurückgreifen kann.

Bibliographische Informationen stellen ein besonders wichtiges verbindendes Element zwischen den zu erschließenden und ggf. zu edierenden Materialien aus dem Nachlass Niklas Luhmanns dar. Ihre Modellierung, Zusammenführung und Visualisierung verspricht einen vollständigen Überblick über Grundlagen, Rezeption und Verbreitung des Luhmannschen Werks aber auch detaillierte Einblicke in seine Arbeitsweise.

Quellen und Forschungsfragen

In Hinblick auf den Werkkosmos Niklas Luhmanns wurde am Institut für Soziologie in Bielefeld eine Bibliographie aller Publikationen

erstellt, die in digitaler Form finalisiert werden soll. Enthalten sind Monographien, Aufsätze, Rezensionen, Festschriften. Gelistet werden außerdem alle bearbeiteten Neu-Auflagen und Übersetzungen. Die insgesamt etwa 1.900 Datensätze sind stark miteinander vernetzt: Sowohl auf der Ebene von Monographien als auch bei Artikeln wird jeweils auf Nachdrucke, Übersetzungen und weitere Auflagen hingewiesen. Die Aufbereitung dieser Informationen soll anschaulich machen, in welchen Medien Luhmann hauptsächlich publiziert hat und wie sich seine Schriften durch Übersetzungen und Nachdrucke international verbreitet haben.



Schematische Darstellung zur Vernetzung der Datensätze

Den Lesekosmos Niklas Luhmanns darstellbar und damit erforschbar zu machen, ist ein zweites zentrales Anliegen. Die folgenden Quellen bibliographischer Information liegen dazu vor:

- die zitierte Literatur in den Zettelkästen
- eigenständige bibliographische Abteilungen je Zettelkasten (ca. 17.800 Einträge)
- die zitierte Literatur in seinen eigenen publizierten Werken
- die Literaturhinweise in seinen unpublizierten Manuskripten
- die private Arbeitsbibliothek (ca. 4.000)

Nach Aufarbeitung dieser Materialien können wichtige werkhistorische Fragestellungen untersucht werden: Welche und wie viel Literatur (aus welchen wissenschaftlichen Disziplinen) hat Luhmann im Laufe seiner Forschung rezipiert bzw. verarbeitet? Gibt es bestimmte Werke / Autoren, die er wiederkehrend konsultiert bzw. zitiert hat? Gibt es bevorzugte Zeitschriften?

Der Zettelkasten ist der zentrale Startpunkt für die Reise in den bibliographischen Kosmos. Er enthält durchgehend auf den Notizzetteln Hinweise auf benutzte Literatur und ab der Mitte der 70er Jahre auch eine eigenständige Bibliographie-Abteilung (insgesamt etwa 17.800 Einträge).

Eine Auswertung der Zitatnachweise in den gedruckten Bänden Luhmanns soll dann die Verbindung von Zettelkasten und Publikationen erleichtern. Wieviel der im Zettelkasten nachgewiesenen Lektürearbeit ist tatsächlich in die Publikationen eingegangen? Und gibt es Stellen in den Publikationen, die scheinbar keine Wurzeln im Zettelkasten haben?

In den unveröffentlichten Manuskripten finden sich neben bereits standardisierten Literaturnachweisen häufig auch handschriftliche Nennungen von Notizzetteln, die den Kreis über den Zettelkasten wieder schließen.

Datenmodell

Zu unseren Zielen gehört die Vernetzung der Datensätze zu visualisieren, um so den Werk- bzw. Lesekosmos erfahrbar zu machen. In Anlehnung an das FRBR-Modell [2] verstehen wir die hier zu modellierenden bibliografischen Datensätze als *Manifestationen* eines *Werks* (Wiesenmüller 2008: 350 **)**. Im Fall der Publikationen Niklas Luhmanns selbst, sind bereits Bezugsinformationen zu anderen *Manifestationen* (z.B. im Sinne von "ist Teil von") und *Expressions* (z.B. im Sinne von "ist Übersetzung von") enthalten. Perspektivisch werden hier noch übergeordnete Datensätze zu *Werken* und deren *Expressions* entstehen, mit denen die Fachwissenschaftler das Material weiter strukturieren und die Werkgenese genauer beschreiben können. [3] Vor allem im Fall der unterschiedlichen Manuskriptfassungen Luhmanns wird die Bestimmung übergreifender *Werke* eine explizit fachwissenschaftliche Aufgabe sein.

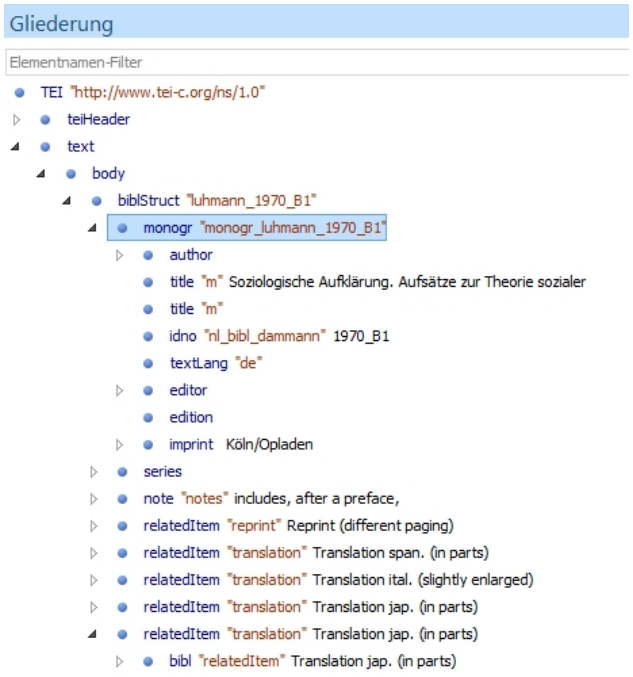
Die Basiskodierung erfolgt zunächst ohne weiterführende ontologischen Differenzierungen und Deutungen in voneinander getrennten Datensätzen auf gleicher Ebene . Die TEI bietet hier mit <biblStruct> ein geeignetes standardisiertes Datenmodell um Manifestationen zu beschreiben, aber auch bereits etablierte Wege wie Personen, Körperschaften, Werktitel und später auch Sachschlagwörter mit Normdatensätzen [4] verbunden werden können. [5] Beziehungen zu anderen Datensätzen werden über <relatedItems> angereichert. Alle Erwähnungen und Verweise auf das bibliograhierte Item werden über <idno>-Elemente aufgenommen . Selbst weiterführende inhaltliche Informationen aus den unterschiedlichen Nachlasszusammenhängen können über <note> mitgeführt werden.

Die Offenheit einer TEI-Auszeichnung ermöglicht uns, im Gegensatz zu anderen bestehenden bibliografischen Formaten, die meist für Bibliothekszusammenhänge zugespitzt wurden , alle vorliegenden Informationen im Datensatz selbst mitzuführen und schwellenlos projektintern für die restlichen in TEI kodierten Materialien nutzbar vorzuhalten . [6]

Umsetzung im Detail

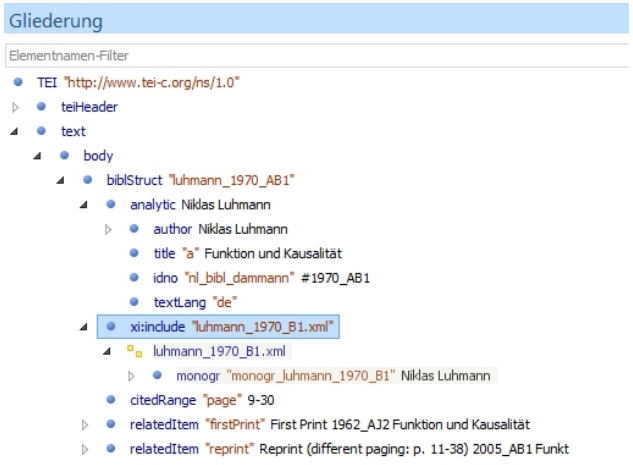
Die Guidelines empfehlen für strukturierte bibliographische Informationen <biblStruct>-Items in Listen (<listBibl>). [7] Wir weichen in diesem Punkt ab und erzeugen für jede Manifestation einen eigenständigen bibliographischen Datensatz in Form eines <biblStruct>-Single-Files. Jeder Datensatz enthält nur ein <monogr> bzw. ein <analytic> und ein <monogr>. Die Dateien erhalten einen eindeutigen Dateinamen (Name des Autors + Erscheinungsjahr + ggf. Erweiterung) und eine entsprechende xml:id. Jedes Vorkommen, etwa im Zettelkasten oder in einem Manuskript, wird in einem <idno>-Element dokumentiert. Hinweise auf Reprints, Übersetzungen und weiterführende Informationen werden in Form von <relatedItems> ergänzt. Da die Dateinamen und xml:ids auf Basis des bibliographierten Werks sprechend benannt wurden, ist eine direkte Verlinkung der Datensätze untereinander problemlos möglich.

Die Aufspaltung in <biblStruct>-Single-Files lässt sich nun auch für die Vorhaltung von unselbstständigen Titeln nutzen. Die Titelinformation für einen Aufsatz wird in <analytic> erfasst, die Information zum Sammelband in einem anschließenden <monogr>-Element. Um die Wiederholung dieser Information für jeden Aufsatz des Sammelbands zu umgehen, wird ein eigenes <biblStruct>-Single-File für den Sammelband erzeugt. Über einen XInclude [8] -basierten Weg wird das <monogr>-Element des Sammelbands in die Datei des Artikels eingebunden. Damit ist die TEI-Datei des Artikels auch während der Bearbeitung vollständig (zusätzlich zu <analytic> wird das externe <monogr> des Sammelbands eingebunden). [9]



Gliederungsansicht eines <biblStruct>-Elements, Typ Sammelband

„



Gliederungsansicht eines <biblStruct>-Elements, Typ “Artikel in Sammelband

Arbeitsumgebung

Als Arbeitsumgebung zur Bearbeitung und Neuerfassung von bibliographischen Informationen kommt ein speziell für dieses Projekt entwickeltes oXygen-Framework [10] zum Einsatz. Ein solches Framework ist eine Erweiterung für den oXygen XML-Editor, welches spezifische Vorgaben für die grafische Darstellung und Funktionsweise eines

Eingabeformulars für das hier entwickelte Datenmodell in oXygen macht. Außerdem werden darin benutzerdefinierte Schaltflächen angelegt, die auf den Workflow des Bearbeiters ausgerichtet sind. Damit ist es für Laien ohne Vorkenntnisse in XML oder TEI auf einfache Weise möglich, bibliographische Informationen auf Grundlage des verwendeten Datenmodells zu erstellen, zu bearbeiten und auszuzeichnen. Das Framework-Verzeichnis wird auf den Rechnern der Bearbeiter zur Verfügung gestellt, von oXygen als solches erkannt und ist damit vom Bearbeiter verwendbar.

oXygen-Framework (Author-Mode)

Veröffentlichung

Nachdem die Daten in die Datenbank importiert wurden, werden sie automatisiert im Projektportal veröffentlicht. Dies geschieht mit einem modular aufgebauten Web-Präsentationssystem, bestehend aus etablierten Open-Source-Softwarelösungen wie eXist XML Database [11] , NodeJS [12] , ReactJS [13] , sowie dem Design-Framework Material Design Lite [14] . Die Datenbank verknüpft automatisch die verschiedenen Datensätze und gibt sie aus, sodass der Benutzer des Portals sofort sehen kann, in welchem Verhältnis ein Werk zu verwandten Werken steht. Eine Visualisierung mittels eines Netzwerk-Graphs soll diese Verknüpfungen zusätzlich veranschaulichen.

Die hier dargestellten Workflows setzen ausschließlich auf Open-Source-Softwarelösungen, sowie offene Standards wie TEI. Die Weitergabe des Frameworks mit allen Templates, der ODD, des Schemas und einer Dokumentation wird angestrebt. Die generische Architektur ist nachhaltig und nachnutzbar von anderen Projekten mit ähnlichen Anforderungen.

Das Luhmann-Projekt eignet sich aufgrund der Heterogenität des bibliographischen Materials sehr gut als Ausgangspunkt zur Entwicklung eines allgemeinen Modells, das vom konkreten Projekt abstrahiert werden kann und soll. Im CCeH wird der Workflow schon von weiteren Projekten eingesetzt und auf seine Tauglichkeit geprüft.

Die für die Publikationen Luhmanns, ihre Reprints und Veröffentlichungen vergebenen Namen und IDs werden, neben den luhmann-basierten Zettelkennungen, als autoritative Identifikatoren für das Werk Luhmanns nachnutzbar sein.

Fußnoten

1. Website des Niklas Luhmann-Archivs http://www.uni-bielefeld.de/soz/luhmann-archiv/ [letzter Zugriff 30 . November 2016]
2. “Functional Requirements for Bibliographic Records” (FRBR) : http://www.ifla.org/publications/functional-requirements-for-bibliographic-records [letzter Zugriff 30 . November 2016]
3. Ein ähnliche Ansatz findet Anwendung im Projekt "Wome n Writer s in Review", vgl http://wwp.neu.edu/review/about/terms [letzter Zugriff 30 . November 2016]
4. Gemeinsame Normdatei der Deutschen Nationalbibliothek (GND), http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html [letzter Zugriff 30 . November 2016]
5. Nach dem FRBR-Modell betrifft das die Entitäten der Gruppe 2 und 3 (Tillet 2010: 3)
6. Bei späteren Exporten in andere bibliographische Formate, wie etwa BibTeX, können Hauptfelder gemappt werden, wohingegen Zusatzinformationen - je nach inhaltlicher Zielsetzung des Ausgabeformat s - schlicht nicht exportiert werden.
7. Vgl. TEI-Guidelines, Abschnitt 3.11: Bibliographic Citations and References (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#COBI [letzter Zugriff 30 . November 2016]
8. Vgl. W3C Empfehlung: https://www.w3.org/TR/xinclude/ und TEI Guidelines http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html#SG-mult [letzter Zugriff 25. August 2016]

9. Für die Weitergabe der Daten nach außenwird das XInclude wieder aufgelöst und für die Verlinkung ergänzte IDs werden gelöscht, so dass die Ausgabe von standardkonformen und vollständigen <biblStruct>-Files sichergestellt ist.
10. Vgl. oXygen Visual (WYSIWYG) XML Editors: https://www.oxygenxml.com/xml_author/WYSIWYG_Editors.html [letzter Zugriff 25. August 2016]
11. Vgl. http://exist-db.org/exist/apps/homepage/index.html [letzter Zugriff 25. August 2016]
12. Vgl. https://nodejs.org/en/ [letzter Zugriff 25. August 2016]
13. Vgl. https://facebook.github.io/react/ [letzter Zugriff 25. August 2016]
14. Vgl. https://getmdl.io/ [letzter Zugriff 25. August 2016]

Bibliographie

Deutsche Nationalbibliothek (Hg.) (2009): *Funktionale Anforderungen an bibliografische Datensätze*. Abschlussbericht der IFLA Study Group on the Functional Requirements for Bibliographic Records. Geänderte und korrigierte Fassung, Februar 2009 (Leipzig/Frankfurt am Main/Berlin, 2009), http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2009_de.pdf . [letzter Zugriff 30. November 2016].

Tillett, Barbara (2003): „What is FRBR?“. Library of Congress Cataloging Distribution Service, 2004 , in: *Technicalities* 25 (5) https://web.archive.org/web/20091229040757/http://www.loc.gov/cds/downloads/FRBR.PDF [letzter Zugriff 30. November 2016].

Wiesenmüller, Heidrun / **Horny, Silke** (2015): *Basiswissen RDA: Eine Einführung für deutschsprachige Anwender*. De Gruyter Saur.

Wiesenmüller, Heidrun (2008): „Zehn Jahre ‚Functional Requirements for Bibliographic Records“: Vision, Theorie und praktische Anwendung“, in: *Bibliothek, Forschung und Praxis* 32 (3).

Niklas-Luhmann-Archiv: Webseite http://www.uni-bielefeld.de/soz/luhmann-archiv/ [letzter Zugriff 30. November 2016].

TEI: *Bibliographic Citations and References in den TEI-Guidelines* http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#COBI [letzter Zugriff 30. November 2016].

IFLA: *Functional Requirements for Bibliographic Records (FRBR)* http://www.ifla.org/publications/functional-requirements-for-bibliographic-records [letzter Zugriff 30. November 2016].

Perspektiven der Benutzeraktionsanalyse im Kontext der Evaluation von Forschungspraktiken in den Digital Humanities

Walkowski, Niels-Oliver
walkowski@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Ansätze zur Evaluation von Forschungsaktivität in den Digital Humanities

Die Dokumentation digitaler Forschungsprozesse ist seit langem ein viel diskutiertes Thema und wird im Allgemeinen mit dem Begriff Provenienz verbunden. Ziel der Erhebung von Provenienz Daten in digitaler Forschung ist es meist den Herstellungsprozess von Ergebnissen aufzuzeichnen um diese dadurch wissenschaftlich nachvollziehbar und reproduzierbar zu machen.

Seit ca. 2-3 Jahren entsteht in den Digital Humanities eine besondere Variante dieses Themas. Konkret geht es dabei um das Modellieren und Dokumentieren von Forschungsprozessen zur Identifikation digitaler geisteswissenschaftlicher Forschungspraktiken, bzw. -methoden. Zu diesem Zweck wurden mit dem *Scholarly Domain Model* (SDM) (Gradmann et al. 2015) und der *NeDiMAH Method Ontology* (NeMO) (Constantopoulos, Dallas, und Bernadou 2016) zwei Modelle geschaffen, die sowohl Konzepte für die Beschreibung von Forschungsprozessen als auch für deren Auswertung in Hinblick auf methodische Fragestellungen bieten.

Der Hintergrund für diese Aktivitäten ist zumeist das Bedürfnis von Infrastrukturprojekten, im Falle der genannten Modelle Europeana und Dariah, Anforderungen von NutzernInnen zu identifizieren und den qualitativen Gebrauch der bereitgestellten Dienste zu evaluieren. Darüber hinaus können sie aber auch als ein digitales Angebot zur Bearbeitung von Fragen der Science Studies

verstanden werden. Im Kontext der Digital Humanities geht diese Perspektive mit dem Wunsch nach der Herausbildung eines methodischen Selbstbewusstseins einher (Constantopoulos, Dallas, und Bernadou 2016).

Versucht man die beiden genannten Ansätze danach zu unterscheiden wie sie sich dem Forschungsprozess nähern, so fällt zunächst ein Unterschied auf. Während in den beispielhaften Anwendungen von NeMO ("DCU OnTo - NeDiMAH Ontology Navigation" 2016) die Beschreibung rückblickend erfolgt, wird sie bei SDM vorausschauend vollzogen. SDM nimmt hier Bezug auf das Konzept des 'modeling for' von Clifford Geertz.

Dieser Unterschied zwingt dazu, die zuvor gewählte Begrifflichkeit noch etwas weiter zu differenzieren. In der Forschungsliteratur werden drei verschiedene Begriff für möglichen Varianten der Dokumentation von Forschungsaktivitäten verwendet (Hunter 2006). Diese sind:

Workflow
Provenienz
Lineage

Die Begriffe können auch als präskriptive, inskribierende, bzw. deskriptive Verfahren bezeichnet werden. Sie unterscheiden sich durch den Zeitpunkt von dem aus die Darstellung eines Forschungsprozesses erfolgt. SDM und NeMO realisieren somit die Workflow und die Lineage Perspektive. Was fehlt ist eine wirkliche Provenienz Perspektive im Kontext der methodischen Evaluation von Forschungsprozessen in den Digital Humanities. Provenienzdaten sind Daten die während einer Aktivität aufgezeichnet werden, indem zum Beispiel bestimmte Aktionen als Trigger für das Abspeichern von Informationen über diese Aktionen dienen.

Der Vorteil einer solchen Vorgehensweise gegenüber den anderen Ansätzen bei der Methodenevaluation ist zweierlei. Die Granularität mit der ein Forschungsprozess beschrieben werden kann ist höher als in SDM und NeMO. Der Anteil inhaltlicher Vorentscheidungen bei der Erfassung der Daten ist geringer. In der Beispielanwendung des SDM soll die Erweiterung der Praxis des Annotierens evaluiert werden. Gleichzeitig nimmt die Entscheidung was wann mit dem im SDM Vokabular bereitgestellten Begriff 'Annotieren' beschrieben wird das Ergebnis der Evaluation schon ein Stück weit vorweg. In einem Provenienz Ansatz reicht es aus zu definieren was aussagekräftige Ereignisse für die Dokumentation des Forschungsprozesses sind. Die Interpretation einer Ereigniskette kann später erfolgen.

Der Vortrag stellt eine beispielhafte Realisierung für ein inskribierendes Verfahren zur Evaluation von Forschungspraxis in den Digital Humanities im zuvor beschriebenen Sinn vor. Die Arbeiten sind Bestandteil der dritten Förderphase des DARIAH-DE Projekts. Der Ausgangspunkt für die Entwicklung des genannten Verfahrens bildet der *Wissensspeicher* ("Digitaler Wissensspeicher" 2016) der *Berlin-Brandenburgischen Akademie der Wissenschaften*.

Der Wissensspeicher ist ein Infrastrukturprojekt, dass eine inhaltliche und technisch Interaktion mit den heterogenen digitalen Ressourcen an der BBAW ermöglicht. Ziel des Wissensspeichers ist es diese Interaktion nicht mit der Objekt-, sondern der hinter den Objekten stehenden Inhaltsebene zu ermöglichen. Die Interaktionen können daher als wissensverarbeitende Prozesse verstanden werden, die es bei entsprechender Evaluation erlauben Strategien der Wissensgenerierung zu identifizierbar.

Verfahren der Benutzer-Interaktionsanalyse im Kontext der Dokumentation von Forschungspraxis

Da die Aufzeichnung der Provenienz Daten die Beschreibung von wissensverarbeitenden Prozessen an Hand von bedeutungsvollen Handlungen zum Ziel hat und nicht einen Transformationsprozesses von Daten, ist die Dokumentation dieser Praxis schwieriger als zum Beispiel bei Hunter und anderen Implementierungen von Provenienz Modellen. Allerdings gibt es einen Forschungsbereich der sich tatsächlich mit einer ähnlichen Fragestellung beschäftigt. Dieser Bereich ist die *Benutzer-Aktions-Analyse*, beziehungsweise *User-Activity-Analysis*. User-Activity-Analysis zeichnet die Mensch-Computer Interaktion mit dem Ziel auf, diese vor dem Hintergrund einer spezifischen Fragestellung zu analysieren. In den meisten Fällen findet User-Activity-Analysis im Browser statt, dies ist jedoch nicht zwingend.

Die zwei Bereiche in denen User-Activity-Analysis hauptsächlich durchgeführt wird sind *e-commerce* und *online-social-networks*. Im Kontext des ersten Bereichs werden

solche Analysen zum Beispiel für den Betrieb von Empfehlungssystemen druchgeführt (Plumbaum, Stelter, und Korth 2009). Bei Online-Social Networks steht häufig die Identifikation von Verhaltensmustern von Menschen in sozialen Interaktionen im Vordergrund (Dang et al. 2016)

Im Kontext akademischer Dienste und Umgebungen ist die User-Activity-Analysis noch nicht so weit verbreitet. Eine Ausnahme bilden spezielle Suchmaschinen für Forschungsliteratur. Beiträge, die dem hier vorgestellten Szenario noch am nächsten kommen sind die von Vozniuk et al. (2016) und Suire et al. (2016), die User-Activity-Analysis im Kontext von *e-learning* und im *cultural heritage* Bereich verorten.

Zwei Aufgabenstellungen sind zunächst einmal konzeptuell von einander zu trennen. Zum einen stellt sich die Frage was, überhaupt als Ausgangspunkt zur Erhebung von Daten in einer spezifischen Mensch-Computer Interaktion dienen kann. Dieser Bereich kann auch als *User-Activity-Tracking* bezeichnet werden. Der andere Bereich umfasst die Frage mittels welchen Verfahren den Ereignissen und auf deren Grundlage gewonnenen Daten Bedeutung zugeschrieben werde kann.

Viele unterschiedliche Strategien der Erhebung von Nutzeraktivitätsdaten sind seit der Entstehung des Web vorgeschlagen worden (Calzarossa, Massari, und Tessera 2016). Die am weitesten verbreitetste ist die sogenannte *Click-Stream-Analyse* bei der die Server-Logdateien, bzw. HTTP-Requests ausgewertet werden, die das Klicken auf Links durch den Benutzer erzeugen. Dieser am einfachsten zu realisierende Ansatz ist aus verschiedenen Gründen problematisch. So enthalten Server-Logdateien keine Informationen über sogenannte 'leise Interaktionen' (Benevenuto et al. 2009), also Klicks, die nicht mit einem Request einhergehen und Interaktionen wie zum Beispiel Mausbewegungen, die gar keinen Klick beinhalten. Darüber hinaus enthalten die erzeugten Daten keinerlei Angaben über den Inhaltskontext in dem die Interaktion stattgefunden hat. Aus diesem Grund wurden Verfahren entwickelt, die auf der Basis von Browser Plug-ins oder mittels JavaScript (Dhawan und Ganapathy 2009) eine detailliertere Dokumentation von Benutzerinteraktion ermöglichen. Ein weiterer Ansatz versucht mittels Parsing- und Miningverfahren Interaktionsspuren oder Inhalte im Kontext von Interaktionen in die Analyse mit einzubeziehen (Vozniuk et al. 2016).

Wesentlich komplexer als die Frage danach welche User-Activity-Daten wie erhoben werden können ist die Frage wie ihnen und den Ereignissen die sie erzeugen Bedeutung beigemessen werden kann. Hier existieren ebenfalls eine Reihe unterschiedlicher Ansätzen. Grundsätzlich lassen sich 5 verschiedene Ansätze unterscheiden. Dazu gehören solche, die die Bedeutung von Ereignissen

- im Vorfeld festlegen,
- vor dem Hintergrund von Durchschnittswerten aus den Daten, die für ein bestimmtes Ereignis erhoben wurden ermitteln
- durch differentielle Verfahren wie zum Beispiel Clusteranalysen ermitteln (Wang et al. 2016)
- unter Einbeziehung des Zustandes der Ereignisumgebung zum Ereigniszeitpunkt wie dem Inhalt der Website bestimmen.
- zu erfassen zu suchen in dem Ereignisse mit externen Quellen wie zum Beispiel Nutzerprofildaten gestellt werden.

Das einzig sinnvolle Verfahren im Kontext des zuvor beschriebenen Ziels ist eine Kombination mehrerer Ansätze sowohl auf der Ebene des Tracking als auch der Analyse. Ausgewertet werden soll wie zuvor umschrieben die Interaktion mit Wissen, also einem Gegenstand, der sich nicht mit einer Trägergröße allein wie z.B. der Website in Übereinstimmung bringen lässt und der per Definition kontextuell konstituiert ist. Wie eine solche Kombination aussehen kann, hängt natürlich von der technischen aber auch sozialen Umgebung ab innerhalb der evaluiert wird.

Eine Architektur zu Evaluation von Forschungsaktivität im Wissensspeicher

Konkret operiert der Use-Case Wissensspeicher im Bereich des Trackings mit einer Kombination aus:

- in das User-Interface hart kodierten expliziten Feedback-Möglichkeiten,
- ins User-Interface integrierte JavaScript snippets, die bei Interaktionen getriggert werden und diese dokumentieren,

- sowie Server-Log Dateien.

Ereignisse werden spezifiziert im Hinblick auf Ereignisgruppen (User Interface, Browser, Request und andere). User Interface Ereignisse werden weiterhin dahingehend kategorisiert auf welchem Seitentyp (Such Interface, Ressourcen Interface und andere) und welchem Layoutbereich sie angesiedelt sind. Schließlich werden für jedes Ereignis variable Eigenschaften bestimmt, die bei der Aufzeichnung dokumentiert werden müssen.

Die Bedeutungsgebung soll ebenfalls durch einen mehrstufigen auf einander aufbauenden Prozess stattfinden. Das Konzept sieht vor, Ereignisse in einem *Task-Model* (Yadav et al. 2015) einzuordnen, welches antizipierte Interaktionsprozesse innerhalb des Wissensspeicher formalisiert. Parallel dazu wird eine erste Auswertung von User-Activity-Daten wiederkehrende Ereignissequenzen identifizieren, die mit dem Task-Model verglichen werden. Dabei auftauchende Fragestellungen lassen sich dann in einem letzten Schritt mittels des *Thinking-Aloud* Verfahrens (Kuusela und Paul 2000) aus dem Bereich des *Usability Testing* bearbeiten. Hierbei arbeitet der Benutzer mit dem entsprechenden Angebot auf dem Computer und beschreibt was er tut, während er es tut.

Der Vortrag wird die vorangegangene Argumentation für den gewählten Ansatz nachzeichnen. Er wird darüber hinaus die ausgewählten Konzepte für die Generierung von User-Activity-Daten und ihrer Bedeutungszuschreibung im Kontext einer methodischen Evaluation von Forschungspraxis im Use-Case diskutieren. Zu guter Letzt wird ein Ausblick auf ein Modell gegeben werden, dass eine gegenseitige Bereicherung zwischen Ansätzen wie SDM und NeMO und dem vorgestellten aufzeigt.

Bibliographie

Benevenuto, Fabrício / Rodrigues, Tiago / Cha, Meeyoung / u. a. (2009): „Characterizing user behavior in online social networks“, in: *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. ACM 49–62.

Calzarossa, Maria Carla / Massari, Luisa / Tessera, Daniele (2016): „Workload Characterization: A Survey Revisited“, in: *ACM Comput. Surv.* 48 (3), 48:1–48:43 10.1145/2856127.

Constantopoulos, Panos / Dallas, Costis / Bernadou, Agiatis (2016): „Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines“, in: Schreibman, Susan / Siemens, Ray / Unsworth, John (edt.): *A New Companion to Digital Humanities*. 1. Aufl. Chichester, West Sussex, UK: John Wiley & Sons.

Dang, Anh / Moh'd, Abidalrahman / Milios, Evangelos / u. a. (2016): „What is in a Rumour: Combined Visual Analysis of Rumour Flow and User Activity“, in: *Proceedings of the 33rd Computer Graphics International*. ACM 17–20.

Dhawan, Mohan / Ganapathy, Vinod (2009): „Analyzing information flow in JavaScript-based browser extensions“, in: *Computer Security Applications Conference, 2009. ACSAC'09. Annual*. IEEE 382–391.

Gradmann, Stefan / Hennicke, Steffen / Tschumpel, Gerold / u. a. (2015): „Beyond Infrastructure! Modelling the Scholarly Domain“.

Hunter, Jane (2006): „Scientific models: a user-oriented approach to the integration of scientific data and digital libraries“, in: *VALA2006* 1–16.

Kuusela, Hannu / Paul, Pallab (2000): „A Comparison of Concurrent and Retrospective Verbal Protocol Analysis“, in: *The American Journal of Psychology* 113 (3): 387–404 10.2307/1423365.

o. A. (o. J.): ***DCU OnTo - NeDIMAH Ontology Navigation***. http://nemo.dcu.gr/ [Letzter Zugriff 26. August 2016].

o. A. (o. J.): ***Digitaler Wissensspeicher***. http://wissensspeicher.bbaw.de/ [Letzter Zugriff 26. August 2016].

Plumbaum, Till / Stelter, Tino / Korth, Alexander (2009): „Semantic Web Usage Mining: Using Semantics to Understand User Intentions“, in: Houben, Geert-Jan / McCalla, Gord / Pianesi, Fabio / u. a. (edt.): *User Modeling, Adaptation, and Personalization*. Berlin / Heidelberg: Springer (Lecture Notes in Computer Science) 391–396.

Suire, Cyrille / Jean-Caurant, Axel / Courboulay, Vincent / u. a. (2016): „User Activity Characterization in a Cultural Heritage Digital Library System“, in: *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM 257–258.

Vozniuk, Andrii / Rodríguez-Triana, María Jesús / Holzer, Adrian / u. a. (2016): „Combining Content Analytics and Activity Tracking to Identify User Interests and Enable Knowledge Discovery“, in: *Proceedings of the 6th Workshop on Personalization Approaches in Learning Environments (PALE 2016)*. 24th conference on User Modeling, Adaptation, and Personalization (UMAP 2016), CEUR workshop proceedings.

Wang, Gang / Zhang, Xinyi / Tang, Shiliang / u. a. (2016): „Unsupervised Clickstream Clustering For User Behavior Analysis“, in:

SIGCHI Conference on Human Factors in Computing Systems.

Yadav, Piyush / Kalyani, Kejul / Mahamuni, Ravi (2015): „User Intention Mining: A Non-intrusive Approach to Track User Activities for Web Application“, in: Tan, Ying / Shi, Yuhui / Buarque, Fernando / u. a. (edt.): *Advances in Swarm and Computational Intelligence*. Springer International Publishing (Lecture Notes in Computer Science) 147–154.

Projekte und Aktivitäten im Kontext digitaler 3D-Rekonstruktion im deutschsprachigen Raum

Münster, Sander
sander.muenster@tu-dresden.de
Medienzentrum / TU Dresden, Deutschland

Kuroczyński, Piotr
piotr.kuroczynski@herder-institut.de
Herder Institut Marburg, Deutschland

Pfarr-Harfst, Mieke
pfarr@dg.tu-darmstadt.de
Fachbereich Architektur, Digitales Gestalten / TU Darmstadt, Deutschland

Einleitung

Die Arbeitsgruppe „Digitale Rekonstruktion“ ging aus der 1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (25.-28.03.2014, Universität Passau) hervor. Die Arbeitsgruppe versammelt Kolleginnen und Kollegen, die sich dem Thema digitale Rekonstruktion aus dem Blickwinkel der Architektur, Archäologie, Bau- und Kunstgeschichte sowie Computergraphik und Informatik verschrieben haben. Die Arbeitsgruppe bietet eine Plattform für einen Austausch und eine feste Etablierung der digitalen Rekonstruktion im Dienste einer Erfassung, Erforschung und Vermittlung kultureller und geschichtlicher Inhalte innerhalb der Digital Humanities. Vorrangiges Ziel der Arbeitsgruppe ist es, die Akteure im deutschsprachigen Raum zusammenzubringen, um sich den Fragen der Begriffsklärung und der Arbeitsmethodik sowie der Dokumentation und Langzeitarchivierung von digitalen Rekonstruktionsprojekten zu widmen. Der Arbeitsgemeinschaft gehören ca. 50 Personen aus 29 Einrichtungen im deutschsprachigen Raum sowie 4 assoziierten Mitgliedseinrichtungen im europäischen Raum an (Abb. 1).

Abb. 1. Example of research which includes the use of digital methods: network analysis as an archival tool

Auf bisher drei Panels wurden bei vergangenen DHd-Jahrestagungen durch die Arbeitsgemeinschaft „Allgemeine Standards, Methodik und Dokumentation“ (Kuroczyński et al., 2014) und „aktuelle Herausforderungen“ (Kuroczyński et al., 2015) sowie Transformationsprozesse „vom digitalen 3D-Datensatz zum wissenschaftlichen Informationsmodell“ (Kuroczyński et al., 2016) beleuchtet. Während damit vor allem Anforderungen und Perspektiven digitaler Rekonstruktion als Forschungs- und Vermittlungsmethode aus generalisierender Perspektive aufgezeigt wurden, sollen in konsequenter Folge auf der DHd 2017 die vielfältigen aktuellen Aktivitäten und Projekte vorgestellt werden, welche sich vor allem im deutschsprachigen Raum mit digitaler 3D-Rekonstruktion in wissenschaftlichen Kontexten beschäftigen. Diese Aufstellung soll dabei eine Bandbreite aktueller Arbeitsschwerpunkte sowohl aufzeigen als auch systematisieren.

Datenbasis

Die Erstellung der Übersicht zu aktuellen Vorhaben und Forschungsperspektiven stützt sich auf drei Zugänge:

1. Eine Aufstellung von Forschungsperspektiven und -vorhaben jenseits des spezifischen Objektes hat das derzeit in Vorbereitung befindliche Buchprojekt der AG Digitale Rekonstruktion unter dem Arbeitstitel „Die Tugend der Modelle 2.0“ zum Ziel. Im Zuge eines Call for Abstracts wurden dabei ca. 20 Beiträge eingereicht, welche nicht nur eine Vielzahl aktueller Projekte beschreiben, sondern auch einen Querschnitt von Forschungskontexten und einer analytische Auseinandersetzung mit dem vielfältigen Themenkomplexen der Digitalen Rekonstruktion widerspiegeln.
2. Bei einer Postersession zum Arbeitstreffen der Arbeitsgemeinschaft im September 2016 werden aktuelle Projekte der Mitglieder der AG vorgestellt. Diese werden ebenfalls in den hier vorgestellten Vortrag Eingang finden und die wissenschaftliche Vielschichtigkeit dieses Themengebietes in der aktuellen Forschugnslandschaft verdeutlichen
3. Wie im folgenden Abschnitt dargestellt, beschäftigen sich zudem eine ganze Reihe von Forschungsvorhaben sowie Graduierungs- und Netzwerkaktivitäten mit einer Kartierung, Systematisierung und Kontextualisierung von Aktivitäten im Kontext digitaler Rekonstruktion und bieten damit gleichermaßen Übersicht über Einzelvorhaben als auch Ansätze für eine Systematisierung von Vorhaben.

Was sind aktuell Forschungs- und Arbeitsschwerpunkte digitaler Rekonstruktion?

Ausgehend von den bereits getätigten Erhebungen lässt sich eine Reihe von Arbeitsschwerpunkten digitaler Rekonstruktion identifizieren.

Anwendung digitaler Rekonstruktion

Der noch immer wesentlichste Kontext einer Anwendung digitaler Rekonstruktion ist die Erstellung digitaler 3D-Modelle konkreter kulturhistorisch bedeutender Objekte wie Siedlungsstrukturen, Einzelgebäude oder Bauwerksensembles sowie Ausstattungsgegenstände ooder Kultobjekte. Dieses dreidimensionale digitale Abbild/ Modell dient vornehmlich zur Vermittlung, aber auch mehr und mehr zur objektbezogenen Forschung. Eine systematische Kartierung von Vorhaben zur objektbezogenen Anwendung digitaler Rekonstruktionen nimmt beispielsweise das Wiki des Arbeitskreises digitalen Kunstgeschichte vor, welches aktuell 3D-Modelle aus ca. 40 Orten auflistet (Arbeitskreis Digitale Kunstgeschichte). Eine thematisch gegliederte Übersicht pflegt daneben Anna Bentkofska-Kafel für das 3D Visualisation in the Arts network (Bentkowska-Kafel). Mit einer inzwischen reichlich 30-jährigen Geschichte und einer Vielzahl von Einzelaktivitäten sowie Zäsuren stellt die digitale 3D-Rekonstruktion daneben inzwischen selbst Gegenstand historiografischer Betrachtungen dar. Beispielhaft dafür sei auf das Dissertationsvorhaben von Heike Messemer verwiesen, welches sich aus kunsthistorischer Perspektive mit einer Erfassung vor allem kunsthistorisch relevanter 3D-Rekonstruktionen und deren Kontextualisierung (Messemer, 2016) beschäftigt.

Systematisierung und methodische Validierung

Digitale Rekonstruktionen nutzen nicht nur Technologien aus der Informatik zur Bearbeitung geisteswissenschaftlicher Fragestellungen, sondern inkorporieren darüber hinaus eine Vielzahl unterschiedlicher disziplinärer Perspektiven und Verwendungskontexte. Neben der Archäologie sowie verschiedenen Aufgaben des Umgangs mit Kulturerbe als Schwerpunkte der EU-Förderung sind in der deutschen Forschungslandschaft spezifische Szenarien, beispielsweise aus Sicht der Kunst- und Architekturgeschichte, Kulturwissenschaft, Bauforschung sowie Museologie, relevant (Riedel et al., 2011, Burwitz et al., 2012). Vor diesem Hintergrund stehen eine Reihe von Vorhaben zur Erfassung und Systematisierung von Forschungs- und Nutzungsansätzen digitaler Rekonstruktion (Münster and Niebling, 2016, Pfarr-Harfst, Forthcoming) sowie generell zu einer wissenschaftlich-methodischen Validierung (vgl. Münster et al., submitted paper).

Modellierung

Im Mittelpunkt digitaler Rekonstruktionen steht die Erstellung eines 3D-Modells anhand der Interpretation historischer Quellen als auch unter Einbeziehung unterschiedlicher Wissensdomänen. Darüber hinaus finden verschiedenste Arten akquirierter Daten Eingang in derartige Projekte, beispielsweise in Form von Laserscans oder photogrammetrische Rekonstruktionen noch existierender Objektteile oder als Landschaftsmodelle. Eine Modellerstellung erfolgt dabei am Computer mittels primär manuell zu bedienender Modellierungssoftwares. Vor dem Hintergrund eines damit verbundenen Aufwands beschäftigt sich eine Reihe von Projekten mit Ansätzen zur Vereinfachung dieser Prozesse durch Vereinfachung von Modellierungswerkzeugen (Schinko et al., 2016, Snickars, 2016, Havemann et al., 2007) oder Abläufen (Ioannides, 2016). Andere dagegen versuchen den manuellen Modellierungsprozess zu strukturieren, allgemeingültige Vorgehensweisen herauszufiltern und diese in digitale, auf Ontologien basierende Handbücher als Beitrag zur Qualitätssicherung zu transferieren (Pfarr-Harfst and Wefers, Forthcoming).

Wissensrepräsentation

Die Erfassung und Archivierung historischer Quellen unterschiedlicher Gattungen, digitalen Forschungsartefakten und -ergebnissen sowie zugeordneten Meta-, Para- und Kontextdaten steht seit langem im Fokus einer Vielzahl von europäischen Vorhaben wie beispielsweise EPOCH, 3D-COFORM, CARARE, 3D-ICONS. Darüber hinaus beschäftigen sich eine Reihe von Arbeiten mit grundlegenden Mechanismen der Dokumentation und Klassifikation digitaler Rekonstruktionen (Pfarr-Harfst, 2013, Huvila, 2014, Münster et al., Forthcoming). Darüber hinaus haben eine Vielzahl aktueller Projekte wie beispielsweise IANUS, Monarch, DocuVis, OpenInfra oder DURAARK die Entwicklung von Forschungsinfrastrukturen zum Ziel (Drewello et al., 2010, Bruschke and Wacker, Forthcoming, Kuroczyński, 2012, Kuroczyński et al., Forthcoming, Beetz et al., Forthcoming). Wenngleich sich diese Vorhaben hinsichtlich des jeweiligen Adressatenkreises und Werkzeugspektrums unterscheiden, werden übergreifend Fragen wie nach Bezüge zwischen Modell und (explizierbaren) Wissensgrundlagen wie beispielsweise Quellen, der Transparentmachung einer Modelllogik (Hoppe, 2001, Günther, 2001), nach dem Modellierungsvorgehen sowie der Beschreibung der erstellten Modelle – beispielsweise mittels übergreifender Referenzontologien und anwendungsspezifischer Applikationsontologien (Homann, 2011, Kuroczyński, 2014) – thematisiert.

Präsentation

Eine Präsentation von 3D-Rekonstruktionen erfolgt schlussendlich wiederum primär in Form von Bildern des erstellten virtuellen 3D-Modells. Mit Blick auf die Qualität dieser Abbildungen ergeben sich besondere Anforderungen dabei hinsichtlich Interaktivität und der Simulationsqualität von Materialität und Lichtstimmung, aber auch im Umgang mit einer heterogenen Belegbarkeit von Hypothesen und zum Umgang mit Alternativhypothesen. Forschungsprojekte beschäftigen sich sowohl mit Fragen der Ästhetik und visuellen Einbeziehung unterschiedlicher Grade von Hypothesenhaftigkeit (Heeb et al., 2016, Vogel, 2016, Lengyel and Toulouse, 2011b, Lengyel and Toulouse, 2011a), als auch mit technologischen Fragen nach Interaktivität und computergrafischer Umsetzung (Fornaro, 2016). Hier schließt sich unmittelbar die Frage nach der Authentizität solcher digitaler Rekonstruktionen an (Pfarr-Harfst, Forthcoming). Mit Blick auf eine Anknüpfbarkeit sind zudem einfach zu bedienende Datenviewer zur Darstellung der 3D-Datensätze relevant, deren Entwicklung beispielsweise im Rahmen der bereits im vorherigen Abschnitt benannten Infrastrukturvorhaben adressiert ist. Ein vergleichsweise neues Präsentationsmedium stellen darüber hinaus 3D-Reproduktion dar (Grellert, Forthcoming), welche virtuelle Modelle in eine Materialität überführen und sich als hybride Präsentationsformen mit den bisher etablierten kombinieren lassen. Wahrnehmung, Didaktik und Präsentation im musealen Kontext sind weitere aktuelle Forschungsthemen(Grellert and Pfarr-Harfst, 2014).

Kompetenzentwicklung

Gerade im geisteswissenschaftlichen Umfeld sind Affinität und Kompetenz hinsichtlich digitaler Forschungsmethoden bisher wenig ausgeprägt (Albrecht, 2013). Ähnlich wie für die Digital Humanities insgesamt (Vorstand des Verbandes Digital

Humanities im deutschsprachigen Raum, 2014) stellt der methodenbezogene Wissens- und Kompetenzaufbau bei Forschern und Praxisanwendern wie bspw. Kuratoren hinsichtlich einer Herstellung, Bewertung und Nutzung digitaler Rekonstruktionen eine wesentliche Herausforderung dar. Entsprechend haben aktuell eine ganze Reihe von Projekten und Netzwerken beispielsweise den Kompetenzerwerb zur Durchführung von Digitalen Rekonstruktionen (Ioannides, 2013, Kröber and Münster, 2016) oder zur Nutzung von Digitalen Rekonstruktionen zur Wissensvermittlung (bspw. Sprünker, 2013, Glaser et al., 2015) zum Ziel.

Vernetzungsaktivitäten

Aktuell umfasst eine Landschaft der digitalen Rekonstruktion in Deutschland eine Vielzahl von Akteuren unterschiedlicher Hintergründe, welche bisher nur ungenügend vernetzt und organisiert sind. Daraus leiten sich der Bedarf gemeinsamer Plattformen für einen Austausch und eine Etablierung der digitalen Rekonstruktion im Kanon der Digital Humanities ebenso wie nach disziplinübergreifenden koordinierende Strukturen bzw. Institutionen einer wissenschaftlichen und anwendungspraktischen Weiterentwicklung ab. Ein diesbezüglich erster Schritt war nicht zuletzt die 2014 erfolgte Gründung der AG „Digitale Rekonstruktion“ der DHd, welche auf europäischer Ebene wiederum in eine Reihe multinationaler und zumeist thematisch begrenzter Netzwerke, beispielsweise zu virtuellen Museen oder Farbe und Raum von Kulturgut (Boochs et al., 2014), eingebunden ist.

Bibliographie

H2020 Virtual Multimodal Museum [Online]. http://www.vi-mm.eu [letzter Zugriff 25. August 2016].

IANUS - Forschungsdatenzentrum Archäologie & Altertumswissenschaften [Online]. http://www.dainst.org/de/project/ianus-forschungsdatenzentrum-arch%C3%A4ologie-altertumswissenschaften?ft=all [letzter Zugriff 25. August 2016].

OpenInfRA - Ein webbasiertes Informationssystem zur Dokumentation und Publikation archäologischer Forschungsprojekte [Online]. http://www.tu-cottbus.de/projekte/de/openinfra/ [letzter Zugriff 25. August 2016].

Albrecht, Steffen (2013): „Scholars' Adoption of E-Science Practices: (Preliminary) Results from a Qualitative Study of Network and Other Influencing Factors“, in: *XXXIII. Sunbelt Social Networks Conference of the International Network for Social Network Analysis (INSNA)*.

Arbeitskreis Digitale Kunstgeschichte: *Liste digitaler Modelle historischer Architektur* [Online]. http://www.digitale-kunstgeschichte.de/wiki/Liste_digitaler_Modelle_historischer_Architektur [letzter Zugriff 25. August 2016].

Beetz, Jakob / Blümel, Ina / Dietze, Stefan / Fetahui, Besnik / Gadiraju, Ujwal / Hecher, Martin / Krijnen, Thomas / Lindlar, Michelle / Tamke, Martin / Wessel, Raoul / Yu, Ran (2016): „Enrichment and Preservation of Architectural Knowledge“, in: Münster, Sander / Pfarr-Harfst, Mieke / Kuroczyński, Piotr / Ioannides, Marinos (eds.): *How to manage data and knowledge related to interpretative digital 3D reconstructions of Cultural Heritage?* Cham: Springer.

Bentkowska-Kafel, Anna: *3DVisA Index of 3D Projects* [Online]. http://3dvisa.cch.kcl.ac.uk/projectlist.html [letzter Zugriff 25. August 2016].

Boochs, Frank / Bentkowska-Kafel, Anna / Degrigny, Christian / Karaszewski, Maciej / Karmacharya, Ashish / Kato, Zoltan / Picollo, Marcello / Sitnik, Robert / Trémeau, Alain / Tsiafaki, Despoina / Tamas, Levente (2014): „Colour and Space in Cultural Heritage: Key Questions in 3D Optical Documentation of Material Culture for Conservation, Study and Preservation“, in: Ioannides, Marinos / Magnenat-Thalmann, Nadia / Fink, Eleanor / Žarnić, Roko / Yen, Alex-Yianing / Quak, Ewald (eds.): *Digital Heritage: Progress in Cultural Heritage: Documentation, Preservation, and Protection5th International Conference, EuroMed 2014: Proceedings*. Cham: Springer.

Bruschke, Jonas / Wacker, Markus (Forthcoming): „Simplifying the documentation of digital reconstruction processes: Introducing an interactive documentation system“, in: Münster, Sander / Pfarr-Harfst, Mieke / Kuroczyński, Piotr / Ioannides, Marinos (eds.): *How to manage data and knowledge related to interpretative digital 3D reconstructions of Cultural Heritage?* Cham: Springer LNCS.

Burwitz, Henning / Henze, Frank / Riedel, Alexandra (2012): „Alles 3D? – Über die Nutzung aktueller Aufnahmetechnik in der archäologischen Bauforschung“, in: Faulstich, Elisabeth Ida (ed.): *Dokumentation und Innovation bei der Erfassung von Kulturgütern*

II, Schriften des Bundesverbands freiberuflicher Kulturwissenschaftler 5, Online-Publikation der BfK-Fachtagung 2012. Würzburg.

Drewello, Rainer / Freitag, Burkhard / Schlieder, Christoph (2010): „Neues Werkzeug für alte Gemäuer", in: *DFG Forschung Magazin* 3: 10–14.

Fornaro, P. (2016): „Farbmanagement im 3D Raum", in: *Der Modelle Tugend 2.0*.

Glaser, Manuela / Lengyel, Dominik / Toulouse, Catherine / Schwan, Stephan (2015): „Designing computer based archaeological 3D-reconstructions: How camera zoom influences attentaion", in: Bares, William / Christie, Marc / Ronfard, Remi (eds.) *Proceedings of the Eurographics Workshop on Intelligent Cinematography and Editing EICED 2015*. Goslar.

Grellert, Marc (2016): „Rapid Prototyping in the Context of Cultural Heritage and Museum Displays. Buildings, Cities, Landscapes, Illuminated Models", in: Münster, Sander / Pfarr-Harfst, Mieke / Kuroczyński, Piotr / Ioannides, Marinos (eds.): *How to manage data and knowledge related to interpretative digital 3D reconstructions of Cultural Heritage?* Cham: Springer LNCS.

Grellert, Marc / Pfarr-Harfst, Mieke (2014): „25 Years of Virtual Reconstructions. Project Report of Department Information and Communication Technology in Architetcture at Technische Universität Darmstadt", in: *18th International Conference on Cultural heritage and New Technologies*.

Günther, Hubertus (2001): „Kritische Computer-Visualisierung in der kunsthistorischen Lehre", in: Frings, Marcus (ed.) *Der Modelle Tugend: CAD und die neuen Räume der Kunstgeschichte*. Weimar.

Havemann, Sven / Settgast, Volker / Lancelle, Marcel / Fellner, Dieter W. (2007): *3D-Powerpoint - Towards a Design Tool for Digital Exhibitions of Cultural Artifacts*. Brighton, UK: Eurographics Association.

Heeb, N./ Christen, J./ Rohrer, J. / Lochau, S. (2016): „Strategien zur Vermittlung von Fakt, Hypothese und Fiktion in der digitalen Architektur-Rekonstruktion", in: *Der Modelle Tugend 2.0*.

Hohmann, Georg (2011): „Die Anwendung von Ontologien zur Wissensrepräsentation und -kommunikation im Bereich des kulturellen Erbes", in: Schomburg, Silke / Leggewie, Claus / Lobin, Henning / Puschmann, Cornelius (eds.): *Digitale Wissenschaft - Stand und Entwickung digital vernetzter Forschung in Deutschland*. Köln: HBZ.

Hoppe, Stephan (2001): „Die Fußnoten des Modells", in: Frings, Marcus (ed.) *Der Modelle Tugend. CAD und die neuen Räume der Kunstgeschichte*. Weimar.

Huvila, Isto (2014): *Perspectives to Archaeological Information in the Digital Society*. Uppsala, Institutionen för ABM och författarna.

Ioannides, Marinos (2013): „Initial Training Network for Digital Cultural Heritage: Projecting our Past to the Future", in: *Der Modelle Tugend 2.0*.

Ioannides, Marinos (2016): „Monument Documentation Engineering", in: *Der Modelle Tugend 2.0*.

Kröber, Cindy / Münster, Sander (2016): „Educational App Creation for the Cathedral in Freiberg", in: Spector, J. Michael / Ifenthaler, Dirk / Sampson, Demetrios G. / Isaias, Pedro (eds.): *Competencies, Challenges, and Changes in Teaching, Learning and Educational Leadership in the Digital Age*. Springer.

Kuroczyński, Piotr (2012): „3D-Computer-Rekonstruktion der Baugeschichte Breslaus: Ein Erfahrungsbericht", in: Polnische Akademie der Wissenschaften (ed.): *Jahrbuch des Wissenschaftlichen Zentrums der Polnischen Akademie der Wissenschaften in Wien 3*. Wien.

Kuroczyński, Piotr (2014): „Digital Reconstruction and Virtual Research Environments – A question of documentation standards. Access and Understanding – Networking in the Digital Era", in: *Proceedings of the annual conference of CIDOC*.

Kuroczyński, Piotr / Grellert, Marc / Hauck, O. / Münster, Sander / Pfarr-Harfst, Mieke / Scholz, Martin (2015): „Digitale Rekonstruktion und aktuelle Herausforderungen (Panel)", in: *DHd 2015: Von Daten zu Erkenntnissen*.

Kuroczyński, Piotr / Hauck, Oliver B. / Dworak, Daniel (Forthcoming): „3D models on triple paths - New pathways for documenting and visualising virtual reconstructions", in: Münster, Sander / Pfarr-Harfst, Mieke / Kuroczyński, Piotr / Ioannides, Marinos / Quack, E. (eds.) *The 2nd International Workshop on ICT for the Preservation and Transmission of Intangible Cultural Heritage ‚How to exchange Cultural Heritage 3D objects and knowledge in Digital Libraries?'*. Cham: Springer.

Kuroczyński, Piotr / Pfarr-Harfst, Mieke / Münster, Sander / Hoppe, Stephan / Hauck, Oliver / Blümel, Ina (2016): „Der Modelle Tugend 2.0 – Vom digitalen 3D-Datensatz zum wissenschaftlichen Informationsmodell", in: *DHd 2016: Modellierung - Vernetzung - Visualisierung*.

Kuroczyński, Piotr / Pfarr-Harfst, Mieke / Wacker, Markus / Münster, Sander / Henze, Frank (2014): „Pecha Kucha: Virtuelle Rekonstruktion – Allgemeine Standards, Methodik und Dokumentation (Panel)", in:

DHd 2014: Digital Humanities – methodischer Brückenschlag oder feindliche Übernahme?. Passau.

Lengyel, Dominik / Toulouse, Catherine (2011): „Darstellung von unscharfem Wissen in der Rekonstruktion historischer Bauten“, in: Heine, Katja / Rheidt, Klaus / Henze, Frank / Riedel, Alexandra (eds.): *Von Handaufmaß bis High Tech III. 3D in der historischen Bauforschung* . Darmstadt: Verlag Philipp von Zabern.

Lengyel, Dominik / Toulouse, Catherine (2011): „Ein Stadtmodell von Pergamon - Unschärfe als Methode für Darstellung und Rekonstruktion antiker Architektur“, in: Petersen, Lars / Hoff, Ralf von den (eds.): *Skulpturen in Pergamon – Gymnasion, Heiligtum, Palast*. Freiburg: Archäologische Sammlung der Albert-Ludwigs-Universität Freiburg.

Messemer, Heike (2016): „The Beginnings of Digital Visualization of Historical Architecture in the Academic Field“, in: Hoppe, Stephan / Breitling, Stefan (eds.): *Virtual Palaces, Part II. Lost Palaces and their Afterlife: Virtual Reconstruction between Science and the Media*.

Münster, Sander / Friedrichs, K. / Hegel, Wolfgang (eingereicht): „3D Reconstruction techniques as a Cultural Shift in Art History?“, in: *International Journal of Digital Art History*.

Münster, Sander / Hegel, Wolfgang / Kröber, Cindy (2016): „A classification model for digital reconstruction in context of humanities research“, in: Münster, Sander / Pfarr-Harfst, Mieke / Kuroczyński, Piotr / Ioannides, Marinos (eds.): *How to manage data and knowledge related to interpretative digital 3D reconstructions of Cultural Heritage?* Cham: Springer LNCS.

Münster, Sander / Niebling, Florian (2016): „Building a wiki resource on visual knowledge related knowledge assets“, in: Spender, J.C. / Schiuma, Giovanni / Nönnig, Jörg Rainer (eds.): *Proceedings of the 11th International Forum on Knowledge Asset Dynamics (IFKAD 2016)*. Dresden.

Pfarr-Harfst, Mieke (2013): *Documentation system for digital reconstructions Reference to the Mausoleum of the Tang-Dynastie at Zhaoling, in Shaanxi Province, China* (unveröffentlicht).

Pfarr-Harfst, Mieke (Forthcoming): „Typical Workflows, Documentation Approaches and Principles of 3D Digital Reconstruction of Cultural Heritage“, in: Münster, Sander / Pfarr-Harfst, Mieke / Kuroczyński, Piotr / Ioannides, Marinos (eds.): *How to manage data and knowledge related to interpretative digital 3D reconstructions of Cultural Heritage?* Cham: Springer LNCS.

Pfarr-Harfst, Mieke / Wefers, Stefanie (2016): „Digital 3D reconstructed models – Structuring visualisation project workflows“, in: Ioannides, Marinos (ed.): *Proceedings of the 6th International Conference, EuroMed 2016*. Cham: Springer.

Riedel, Alexandra / Henze, Frank / Marbs, Andreas (2011): „Paradigmenwechsel in der historischen Bauforschung? Ansätze für eine effektive Nutzung von 3D-Informationen“, in: Heine, Katja / Rheidt, Klaus / Henze, Frank / Riedel, Alexandra (eds.): *Von Handaufmaß bis High Tech III - 3D in der historischen Bauforschung*. Darmstadt: Philipp von Zabern.

Schinko, C. / Krispel, U. / Gregor, R. / Schreck, T. / Ullrich, T. (2016): „Generative Modeling – the Combination of Knowledge and Geometry“, in: *Der Modelle Tugend 2.0*.

Snickars, Pelle (2016): „Metamodeling. 3D-(re)designing Polhem’s Laboratorium mechanicum“, in: *Der Modelle Tugend 2.0*.

Sprünker, Janine (2013): „Making online cultural heritage visible for educational proposes“, in: *Digital Heritage International Congress (DigitalHeritage)* 405–408.

Vogel, G.-H. (2016): „Von der Zweidimensionalität zur Dreidimensionalität: wissenschaftliche Rekonstruktion verlorener Architekturen als archäologische und kunsthistorische Wissensbilder vor dem Hintergrund ästhetischer Konzepte der Kunst- und Architekturgeschichte (Draft)“, in: *Der Modelle Tugend 2.0*.

Vorstand Des Verbandes Digital Humanities Im Deutschsprachigen Raum (2014): *Digital Humanities 2020*. Passau.

„Quellen aus der Schweiz für die Welt: jederzeit, überall, für alle“ – Neue Kooperationen der NB im digitalen Zeitalter

von Wartburg, Karin
karin.vonwartburg@nb.admin.ch
Schweizerische Nationalbibliothek, Schweiz

Nepfer, Matthias
Matthias.Nepfer@nb.admin.ch
Schweizerische Nationalbibliothek, Schweiz

Die *Schweizerische Nationalbibliothek NB* ist eine Gedächtnisinstitution des Bundes. Gemeinsam mit anderen Bibliotheken, mit den Archiven und Museen trägt sie zur Erhaltung des kulturellen Erbes der Schweiz bei. Sie überliefert Texte, Bilder und Töne, die einen Bezug zur Schweiz haben, auf analogen und digitalen Trägern. Sie verfügt inzwischen über rund fünf Millionen Dokumente, die zum grössten Teil seit der Gründung des Bundesstaates 1848 entstanden sind. Zu der NB gehören das Schweizerische Literaturarchiv SLA, das Centre Dürrenmatt Neuchâtel CDN und die Fonoteca Nazionale FN.

In ihrer *Strategie* stellt sich die NB den digitalen Herausforderungen und bekennt sich dazu, ihre Inhalte weltweit zugänglich machen zu wollen: „Quellen aus der Schweiz für die Welt." Damit soll jeder und jedem Einzelnen ermöglicht werden, diese Dokumente für die eigenen Bedürfnisse zu nutzen. Im Fokus sind dabei Personen, für die die Sammlung der NB von Bedeutung ist: Studierende, Fachleute und Forschende der Kulturwissenschaften, vor allem aber die Schweizer Bevölkerung. In strategischen Handlungsfeldern wird festgelegt, dass die Sammlung der NB leicht zu finden und einfach zu benutzen sein soll. Ausserdem wird der Anspruch formuliert, die Personen, die an unseren Beständen forschen mit Dienstleistungen und Beratung wirkungsvoll zu unterstützen. Im Fokus sind dabei die Literaturwissenschaft, die Schweizer Geschichte und die Auswertung von Bildbeständen.

Ausgangslage für die Formulierung dieser strategischen Handlungsfelder war eine *Umfeldanalyse*, bei der zwei Herausforderungen und einen Trend identifiziert worden waren.

Eine *erste Herausforderung* besteht darin, dass jede Informationssuche im Internet mit einer Suchmaschine beginnt – nicht mit einem Bibliothekskatalog, einer Archivdatenbank oder einem Portal von Gedächtnisinstitutionen. Dies entspricht der eigenen persönlichen Erfahrung und es wird durch diverse Studien (Eine Zusammenstellung der Befunde bei Silipigni Connaway et al. 2010) bestätigt: Es ist auch bekannt, dass die Online-Enzyklopädie Wikipedia unter den ersten Treffern erscheint, sofern darin ein Artikel zum gesuchten Thema vorhanden ist. Die relevanten Metadaten und/oder Inhalte von Gedächtnisinstitutionen erscheinen – wenn überhaupt – viel weiter unten in der Liste der Suchresultate.

Eine *andere Herausforderung* sind die hohen Erwartungen der Benutzenden: Diese wollen nicht nur die Metadaten, also die Beschreibung von Inhalten finden, sondern auf diese jederzeit und von überall her zugreifen können, um sie sofort für die eigenen Zwecke verwenden zu können.

Der Open- *Trend* ist eine Chance: Der Ruf nach Öffnung von Daten ist ein weltweiter Trend, die Rede ist von Open Access, Open Government Data, Open Data, OpenGLAM, usw. Gedächtnisinstitutionen – oder eben die mit dem Akronym GLAM gemeinten Galleries, Libraries, Archives and Museums – folgen bei der Erschliessung internationalen Standards und verwenden Normdaten. Ihre Metadaten, vor allem die bibliografischen Beschreibungen der Bibliotheken, gelten als qualitativ hoch stehend. Mit den fortwährenden Digitalisierungsbemühungen werden ausserdem laufend attraktive Inhalte auf verschiedenen Plattformen online gestellt. Communities wie die Wikipedianer, Open Data- und Public Domain-Aktivistinnen, sowie Forschende der Digital Humanities interessieren sich für diese Daten, sofern sich diese dank einer freien Lizenz problemlos weiterverwenden lassen.

Die *Umsetzung der Strategie* hat zu neuen Handlungsfeldern und auch zu neuen Kooperationen geführt, die in diesem Vortrag vorgestellt werden sollen.

Neue Kooperationen und Aktivitäten sind beispielsweise im Umfeld von *OpenGLAM* [1] zu verzeichnen.

- Die NB strebt an, die eigenen Daten gemäss den von Open Knowledge International verabschiedeten Prinzipien sichtbar und für diverse Nutzungen möglichst frei verfügbar zu machen. Ausserdem soll die Nachnutzung der offenen Daten aktiv gefördert werden.
- Die NB stellt den Forschenden resp. allen an ihrer Sammlung interessierten Personen Metadaten, Normdaten, Digitalisate (Text und Bilder) und originär digitale Ressourcen (Webseiten, e-Medien) zur Verfügung. Die Daten werden nach Möglichkeit offen, ohne organisatorische, technische oder finanzielle Hürden zur Verfügung gestellt. Zusätzlich zum eigenen Bibliothekssystem und der eigenen Archivdatenbank werden dafür gut sichtbare, stark frequentierte Plattformen wie die Mediendatenbank der Online-Enzyklopädie Wikipedia, Wikimedia

Commons oder das Portal für Schweizer Behördendaten opendata.swiss verwendet.
- Um die Nutzung der Daten zu fördern und mit interessierten Communities in Kontakt zu treten, beherbergte die NB 2015 den ersten Kulturdaten-Hackathon in der Schweiz, an dem auch Forschende der Digital Humanities teilnahmen. Der an diesem Hackathon entwickelte Prototyp Gugelmann-Galaxy zeigt exemplarisch, welche unerwarteten, innovativen Nutzungen „geschehen“ können wenn Gedächtnisinstitutionen gezielt Teile ihrer Sammlung aus den Datensilos befreien und der Öffentlichkeit zur Weiterverwendung überlassen.
- Grundlage für diese OpenGLAM-Aktivitäten waren einerseits die 2012 verabschiedete Open-Data-Strategie der NB, andererseits die 2013 abgeschlossene Kooperationsvereinbarung mit Wikimedia Schweiz. Nachdem die NB beschlossen hatte, ihre Metadaten und Inhalte „möglichst offen“ zur Verfügung zu stellen, war der Weg frei, mit Wikimedia Schweiz eine langfristige Zusammenarbeit zu vereinbaren und als erste Massnahme temporär zwei Wikipedians in Residence (vgl. https://en.wikipedia.org/wiki/Wikipedian_in_residence#cite_note-Outreach-18) zu beherbergen.

Ein weiteres Handlungsfeld ist im Bereich der *digitalen Erschliessung* entstanden. Hier stehen laufendende Kooperationen mit Akteuren der Schweizer Geschichtswissenschaft im Fokus, deren Unterstützung, u.a. durch die Weiterentwicklung der von der NB herausgegebene Bibliographie der Schweizer Geschichte BSG, ein strategisches Ziel darstellt.

- Ein wichtiges Projekt in diesem Bereich ist Metagrid [2] bei welchem die NB seit Beginn des Projekts als Projektpartner beteiligt ist. Der Webservice Metagrid ermöglicht die Einrichtung, Verwaltung und Analyse von Links zwischen identischen Identitäten von verschiedenen Websites und Datenbanken. Seit Sommer 2016 sind der Katalog Helveticat und die BSG in Metagrid integriert und via Metagrid-Widget abrufbar. Zur Zeit beschränkt sich der Webservice auf Personennamen. Die NB liefert dem Webservice Namen von Autoren und Persönlichkeiten, welche mit einer GND-Nummer verknüpft sind. Auf diese Weise können andere Projektpartner von Metagrid die eindeutige, in der Bibliothekswelt weit verbreitete Identifikationsnummer direkt übernehmen.
- In einem Pilotprojekt mit der Rechtsquellenstiftung des Schweizerischen Juristenvereins werden seit 2015 von der Rechtsquellenstiftung für deren Online-Editionen benötigte Literatur in der Datenbank der BSG erfasst, resp. die schon in der BSG vorhandenen Katalogisate angepasst. [3] Dank dieser Kooperation können Doppelspurigkeiten bei der Literaturerfassung vermieden, durch die gegenseitige Verlinkung die Visibilität der verschiedenen Projekte erhöht und die weiterführende Recherche für die Benutzenden vereinfacht werden. Die BSG entwickelt sich dadurch von einem auf blosse bibliografische Nachweise orientierten Literaturverzeichnis hin zu einem „Literaturportal“ oder „Informationsraum“ für Literatur zur Schweizer Geschichte (Wissen 2008: 223ff.). Datenbanken mit historischem Content erhalten so die Möglichkeit, Literaturnachweise zur Schweizer Geschichte in der BSG zu holen, zu verlinken und barrierefrei nachzunutzen.

Im Vortrag werden ausgewählte Resultate präsentiert, die für Forschende der Digital Humanities potentiell von Interesse sein könnten, wie zum Beispiel die Inhalte der NB auf Wikimedia Commons , opendata.swiss oder dem Pilotportal Linked Data Service LINDAS, der Prototyp Gugelmann-Galaxy , Verwendungen des Webservices Metagrid und die Veranstaltung Open Cultural Data Hackathon. Ausserdem wird die Nutzung resp. der Nutzen aus Sicht der NB thematisiert und ein vorläufiges Fazit gezogen. Am Schluss wird bezüglich den Handlungsfeldern und den Kooperationen ein Ausblick in die nähere und fernere Zukunft gewagt.

Fußnoten

1. OpenGLAM ist eine Initiative von Open Knowledge International die eine Öffnung der Gedächtnisinstitutionen propagiert. Das Schweizer Chapter von Open Knowledge International ist der Verein opendata.ch .
2. Metagrid ist ein Projekt der Schweizerischen Akademie der Geisteswissenschaften SAGW, durchgeführt von den Diplomatischen Dokumenten der Schweiz DDS mit der Unterstützung des Historischen Lexikons der Schweiz HLS. Vgl. www.metagrid.ch

3. Im Rechtsquellenportal des Staatsarchivs Zürich werden die in der Datenbank BSG bearbeiteten bibliografischen Aufnahmen im Literaturverzeichnis abgebildet (http://www.rechtsquellen-online.zh.ch/startseite/literaturverzeichnis). In der Personendatenbank der Rechtsquellenstiftung hingegen werden für weiterführende Literatur direkt Links auf die Datenbank BSG gesetzt (https://www.ssrq-sds-fds.ch/persons-db/?query=per001666&query-type=perid).

Bibliographie

Estermann, Beat (2015): „Diffusion of Open Data and Crowdsourcing among Heritage Institutions. Based on data from Finland, Poland, Switzerland, and The Netherlands“, in: *EGPA 2015 Conference* http://survey.openglam.ch/publications/EGPA2015_Estermann_Diffusion_of_Open_Data_and_Crowdsourcing_among_Heritage_Institutions_20150901.pdf [letzter Zugriff 26. August 2016].

Estermann, Beat (2013): *Swiss Heritage Institutions in the Internet Era: Results of a pilot survey on open data and crowdsourcing, 2013*. http://espace.okfn.org/items/show/226 [letzter Zugriff 26. August 2016].

Johnson, Larry / **Adams Becker, Samantha** / **Estrada, Victoria** / **Freeman, Alex** (2015): *NMC Horizon Report: 2015 Library Edition*. Austin, Texas: The New Media Consortium http://www.nmc.org/publication/nmc-horizon-report-2015-library-edition/ [letzter Zugriff 26. August 2016].

Koller, Guido (2016): *Geschichte digital. Historische Welten neu vermessen*. Stuttgart: W. Kohlhammer.

„**GLAM & Wikimedia**“, in: *arbido*, Ausgabe 3, 3. September 2015 http://www.arbido.ch/userdocs/arbidoprint/arbido_2015_3_low.pdf [letzter Zugriff 26. August 2016].

Open Knowledge Foundation (2013): *OpenGLAM Principles*. http://openglam.org/principles/ [letzter Zugriff 26. August 2016].

Sanderhoff, Merete (ed.) (2014): *Sharing is caring. Openness and sharing in the cultural heritage sector*. http://www.smk.dk/en/about-smk/smks-publications/sharing-is-caring/ [letzter Zugriff 26. August 2016].

Pekel, Joris (2014): *Democratising the Rijksmuseum*, Europeana Foundation http://espace.okfn.org/items/show/260 [letzter Zugriff 26. August 2016].

Schweizerische Nationalbibliothek (2014): *Strategie 2012 – 2019*, Version 2014 http://www.nb.admin.ch/org/00779/index.html?lang=de&download=NHzLpZeg7t,lnp6I0NTU042l2Z6ln1acy4Z [letzter Zugriff 26. August 2016].

Silipigni Connaway, Lynn / Dickley, Timothy J. (2010): *The Digital Information Seeker*: Report of the Findings from Selected OCLC, RIN, and JISC User Behaviour Projects http://www.jisc.ac.uk/media/documents/publications/reports/2010/digitalinformationseekerreport.pdf [letzter Zugriff 26. August 2016].

Terrass, Melissa (2016): *Opening Access to Collections: the Making and Using of Open Digitised Cultural Content*. http://espace.okfn.org/items/show/259 [letzter Zugriff 26. August 2016].

The Europeana Public Domain Charter (2010): http://pro.europeana.eu/files/Europeana_Professional/Publications/Public%20Domain%20Charter%20-%20DE.pdf [letzter Zugriff 26. August 2016].

Wissen, Dirk (2008): *Zukunft der Bibliographie – Bibliographie der Zukunft. Eine Expertenbefragung mittels Delphi-Technik in Archiven und Bibliotheken in Deutschland, Österreich und der Schweiz*. Berlin: Logos.

Semantische Suche in Ausgestorbenen Sprachen: Eine Fallstudie für das Hethitische

Daxenberger, Johannes
daxenberger@ukp.informatik.tu-darmstadt.de
Ubiquitous Knowledge Processing Lab, Department of Computer Science, Technische Universität Darmstadt

Görke, Susanne
goerkes@uni-mainz.de
Altorientalische Philologie, Institut für Altertumswissenschaften, Johannes Gutenberg-Universität Mainz

Siahdohoni, Darjush
siahdohoni@googlemail.com
Ubiquitous Knowledge Processing Lab, Department of Computer Science, Technische Universität Darmstadt

Gurevych, Iryna
gurevych@ukp.informatik.tu-darmstadt.de
Ubiquitous Knowledge Processing Lab, Department of Computer Science, Technische Universität Darmstadt

Prechel, Doris
prechel@uni-mainz.de
Altorientalische Philologie, Institut für Altertumswissenschaften, Johannes Gutenberg-Universität Mainz

Einleitung

Mit dem Auftreten der Keilschrift am Ende des 4. Jt. v. Chr. bis zur Zeitenwende sind zahlreiche Sprachen des Vorderen Orients aufgezeichnet, deren Kenntnis sich heute allein dem Erhalt der Schriftträger dankt: Eine nicht mehr überschaubare Anzahl von Tontafeln stellt das wesentliche Medium zur Rekonstruktion einer alle menschlichen Lebensbereiche umfassenden dreitausendjährigen Geschichte der heutigen Staaten Syrien, Libanon, Türkei, Irak und Iran dar. Zu den besser bezeugten Sprachen gehört neben dem semitischen Akkadischen das isolierte Sumerisch und das indoeuropäische Hethitisch. Auch wenn sich inzwischen diverse Projekte mit der Digitalisierung des keilinschriftlichen Kulturschatzes befassen, z.B. Cohen et al. (2004) und Tyndall (2012), ist der Zugang zu den kulturell, historisch und linguistisch hochbedeutsamen Textcorpora, die zu großen Teilen noch unpubliziert in den Museen der Welt lagern, meist auf Fachwissenschaftler begrenzt. Um eine adäquate Verwendung der durch Grabungen stetig wachsenden Anzahl von Texten auch in fernerliegenden Arbeitsbereichen zu ermöglichen, ist ein umfassendes Angebot von Übersetzungen in moderne Sprachen höchst wünschenswert.

Das hier skizzierte Projekt zielt insbesondere auf den Umstand, dass selbst die (wenigen) vorhandenen Übersetzungen aufgrund der Durchdringung mit autochtonen Termini es oft an Verständlichkeit vermissen lassen. Das Ziel unserer Pilotstudie ist eine digitale Annäherung an Keilschriftsprachen. Wir stellen eine erweiterte Suchfunktion vor, die es auch fachfremden Benutzern erlaubt, intelligente Suchanfragen in den hethitischen und akkadischen Textcorpora zu stellen. Dazu verwenden wir moderne Natural Language Processing (NLP) Methodologie, die automatisiert lexikalisch-semantische Informationen in mehrsprachigen Übersetzungen von aktuell gut 500 Keilschriftdokumenten extrahiert. Durch den Einsatz vollautomatischer Methoden ist das Hinzufügen neuer Übersetzungen jederzeit möglich – es gibt alleine für das Akkadische über eine halbe Million (noch) nicht digitalisierter Quelltexte. Das Ergebnis unserer Studie ist in Form eines webbasierten Tools verfügbar und wurde in einer Benutzerstudie evaluiert.

Die primären Anforderungen an das Tool [1] sind a) die Rückgabe von Suchergebnissen, die neben exakten oder fast exakten Treffern auch solche enthalten, die aufgrund semantischer Ähnlichkeit zustande kommen, sowie b) eine intuitive Bedienung durch Nutzer, die weder mit der Sprache noch mit sonstigen kulturellen Gegebenheiten vertraut sind.

Vorarbeiten

Bereits seit Längerem wird an der digitalen Methodik zur Verarbeitung von Sprachen des Alten Orients geforscht. Dabei spielte insbesondere die automatisierte morphologische Verarbeitung eine Rolle, siehe bspw. Barthélemy (1998) und Kataja (1988). Neuere Arbeiten setzen größtenteils auf statistische Verfahren anstelle von regelbasierten Ansätzen. Darunter fallen bspw. Liu et al. (2015) mit einer Studie zur Lemmatisierung für Sumerisch sowie Homburg und Chiarcos (2016) zur Wort-Segmentierung im Akkadischen. Im Rahmen des ORACC Projekts werden Tools zur Annotation der Morphologie in Keilschriftsprachen entwickelt, überwiegend für Akkadisch und Sumerisch. [2] Zur semantischen Analyse von Keilschrifttexten existieren hingegen kaum Arbeiten. Lediglich Jaworski (2008) entwickelte eine Ontologie für sumerische ökonomische Aktivitäten, die mit einer semantischen Grammatik dargestellt werden können. Einen Überblick über die lexikalisch-semantischen Analyseverfahren, die in dieser Arbeit zum Einsatz kommen, gibt bspw. Gurevych et al. (2016). Soweit uns bekannt ist, gab es bislang keine Studien, die untersuchen, inwiefern Keilschrifttexte bzw. deren Übersetzungen mittels semantischer-lexikalischer Verfahren für ein breiteres Publikum zugänglich gemacht werden können.

Methodik

Um semantische Suche in Keilschrifttexten zu ermöglichen, haben wir zunächst die transliterierten und übersetzten Texte

vorverarbeitet und für die Suche indexiert. Danach werden sie in einer Datenbank abgelegt, in der mittels einer webbasierten Oberfläche gesucht werden kann.

Daten

Die Texte, die im Rahmen dieser Studie verarbeitet wurden, sind überwiegend hethitische, in Keilschrift verfasste Dokumente (Wilhelm 2008). Die Transliterationen und Übersetzungen (auf Deutsch, Englisch, Italienisch und Französisch) wurden an der Johannes Gutenberg-Universität Mainz sowie von Partnern an weiteren Forschungseinrichtungen im In- und Ausland erstellt. Die Originaltexte stammen aus Anatolien (heutige Türkei) und datieren in die zweite Hälfte des 2. Jt. v. Chr. Inhaltlich handelt es sich vornehmlich um religiöse Texte wie bspw. Gebete oder Rituale. Die Dokumente sind auf Satz- oder Teilsatzebene übersetzt und mit den Transliterationen abgeglichen, so dass einfach Bezüge zwischen den Übersetzungen und den Transliterationen hergestellt werden können. Für jedes Dokument existiert ein Einleitungstext, sowie jeweils eine (kommentierte) Übersetzung und eine Transliteration, siehe Abbildung 1. Die Texte sind unabhängig von dieser Arbeit online zugänglich. [3]

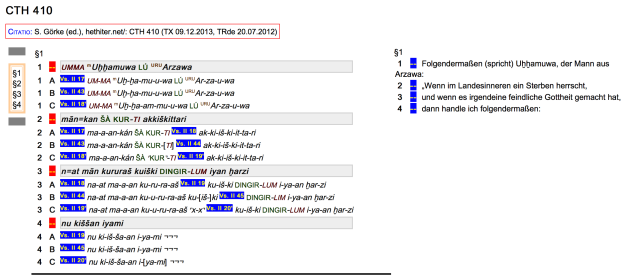


Abbildung 1: Eine manuell erstellte Transliteration (links) und normalisierte Übersetzung (rechts). Quelle: http://www.hethport.uni-wuerzburg.de

NLP Pipeline zur Vorverarbeitung der Texte

Die Übersetzungen und Transliterationen werden direkt aus einem Textformat in eine Pipeline eingelesen, die die weitere linguistische Vorverarbeitung übernimmt. Diese Pipeline erkennt die Struktur der Eingabedokumente, bspw. Dokument-Titel, Sätze, Absätze oder Fußnoten. Außerdem werden die zusammengehörigen Übersetzungen und Transliterationen auf (Teil-)Satzebene gekoppelt. Anschließend werden die mehrsprachigen Übersetzungen mit Hilfe des NLP Frameworks DKPro Core (Eckart de Castilho und Gurevych 2014) analysiert. DKPro Core vereint die Verwendung verschiedener NLP Werkzeuge zur linguistischen Verarbeitung. So ist es möglich, den Inhalt der Dokumente in vier Sprachen zu segmentieren, zu lemmatisieren und nach Wortarten auszuzeichnen. [4] Im nächsten Schritt werden unter Zuhilfenahme des Lesk Algorithmus (Lesk 1986) mehrdeutige Lemmata anhand ihres Kontexts disambiguiert. Dieser Schritt ist die Voraussetzung für die anschließende Zuweisung von sogenannten semantischen Labeln, die einzelne Lemmata mit abstrakteren Konzepten anreichert. Bspw. werden Verben, die eine Bewegung anzeigen, mit einem Label „Bewegung“ gekennzeichnet. Als Ergänzung zu diesen vollautomatischen Verfahren erlaubt es die Pipeline, manuell erstellte Listen für alternative Schreibweisen und Hyperonyme anzuwenden. [5] Darin enthalten sind bspw. geographische Einheiten oder Namen von hethitischen Königen oder Gottheiten, die in den lexikalisch-semantischen Ressourcen, die im Schritt zuvor eingesetzt werden, nicht oder nur teilweise enthalten sind. Bspw. werden verschiedene Namen des Wettergottes (u.a. Taru, Teššup) als solche gelistet. Das Endresultat der Pipeline wird in einem Zwischenformat gespeichert, so dass es anschließend in eine Datenbank importiert werden kann.

Semantische Suchmaschine: Back- und Frontend

Eine MYSQL Datenbank nimmt die Dokumente inklusive der von der NLP Pipeline generierten zusätzlichen semantischen Informationen auf und legt diese in indexierten Tabellen ab. Suchanfragen über das Webinterface werden in entsprechende Abfragen auf die Tabellen übersetzt. Die Anordnung der Suchergebnisse wird über eine Priorisierung der verschiedenen zusätzlichen Informationen geregelt. Wörtliche Treffer werden entsprechend höher gerankt als solche, die durch Übereinstimmung mit semantischen Labeln oder alternativen Schreibweisen zustande kommen.

Das Frontend der Suchmaschine besteht aus dem Eingabefeld für einen oder mehrere Suchbegriffe. Die Suchergebnisse werden pro Dokument gebündelt und angeordnet nach der Güte der Übereinstimmung mit dem Suchbegriff. Abbildung 2 zeigt die Benutzeroberfläche nach einer Suchanfrage. Ein Klick auf ein Suchergebnis öffnet ein Fester, das den Inhalt des gesamten Dokuments jeweils als Übersetzung und Transliteration zeigt.

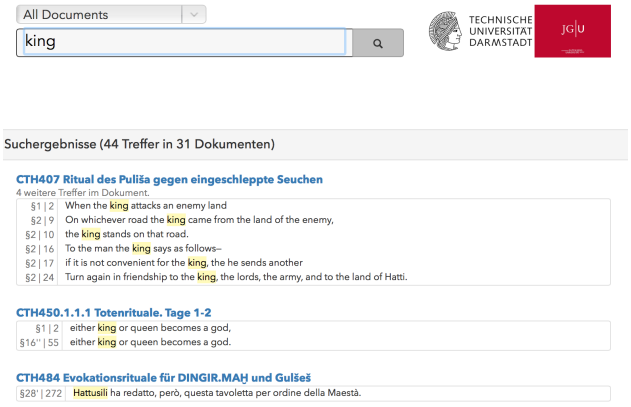


Abbildung 2: Das Frontend mit Ergebnissen zu einer Suchanfrage.

Evaluation

Um zu überprüfen, ob die Suchmaschine die eingangs gestellten Anforderungen erfüllt, haben wir eine anonyme Online-Benutzerstudie mit 23 Fragen unter 27 Teilnehmern [6] durchgeführt. Die Mehrheit der Teilnehmer waren Studierende an deutschen Universitäten. Etwa die Hälfte hatte einen geisteswissenschaftlichen Studienhintergrund, die andere Hälfte einen technischen. Inhaltlich bestand die Benutzerstudie aus einer kurzen Einleitung sowie drei Teilen mit Fragen. Der erste Teil beinhaltete einfache Fragen, die das allgemeine Verständnis der Suchabfragen überprüfen sollten (bspw. Suche nach einem Begriff in einem bestimmten Dokument). Der zweite Teil zielt explizit auf den semantischen Teil der Suchfunktion ab (bspw. Suche nach dem Namen einer Gottheit). Im dritten Teil wurde die allgemeine Bedienbarkeit und Nützlichkeit des Tools erfragt.

Mit wenigen Ausnahmen wurden die Aufgaben aus dem ersten Teil der Benutzerstudie von allen Teilnehmern korrekt gelöst. Im zweiten Teil mussten diverse hethitische Gottheiten, Könige oder Städte namentlich benannt werden, diese Aufgabe konnten sämtliche Teilnehmer korrekt lösen. Eine Frage, in der die (nicht vorhandene) Beziehung zwischen zwei Gottheiten anhand von Suchergebnissen bestimmt werden sollte, wurde nur von etwa zwei Dritteln der Teilnehmer korrekt gelöst. Tabelle 1 fasst die Abfragen und Ergebnisse aus dem dritten Teil der Benutzerstudie zusammen.

Kriterium	Durchschnittswert (1-5)
Sortierung der Ergebnisse basierend auf einer konkreten Suchanfrage	4.15
Benutzerfreundlichkeit der Weboberfläche	4.19
Allgemeine Qualität der Suchergebnisse	4.26
Nützlichkeit für Fachfremde	4.3

Tabelle 1: Kriterien und Bewertungen (Auswahlmöglichkeiten zwischen 1 = sehr schlecht und 5 = sehr gut) des dritten Teils der Benutzerstudie.

Diskussion

In der Gesamtheit zeigen die Ergebnisse der Benutzerstudie, dass das Tool die eingangs gestellten Anforderungen erfüllt. Neben den zu lösenden Aufgaben gab es auch die Möglichkeit, per Freitextfeld Rückmeldung zu geben. Die so identifizierten Probleme sind zurückzuführen auf a) fehlende Erklärungen zur Formulierung von Suchanfragen, b) Fehlern in den manuell erstellten Listen mit alternativen Schreibweisen und Hyperonymen, und c) irreführender Hervorhebung von Wörtern bei Ergebnissen, die auf semantische Übereinstimmung zurückzuführen sind.

Zu den allgemeinen Herausforderungen bei der Aufbereitung der Daten für die semantische Suche zählt u.a. die Fragmenthaftigkeit diverser Texte. Da solche Phänomene zu Fehlern in der NLP Vorverarbeitung (bspw. bei der Segmentierung) führen, wurde eine Komponente in die Pipeline integriert, die Lücken soweit möglich repariert. Der „Vocabulary Gap“ zwischen den Termini in den lexikalisch-semantischen Ressourcen und dem in den Übersetzungen tatsächlich verwendeten Vokabular hat letztlich dazu geführt, dass zusätzlich manuell erstellte Wortlisten eingesetzt wurden. Diese Listen müssen allerdings nur einmal erstellt werden und haben einen überschaubaren Umfang.

Neben der Behebung der oben genannten Probleme ist als nächster Schritte u.a. vorgesehen, das Backend um eine Funktion zum einfachen Upload neuer, transliterierter und übersetzter Texte in die Datenbank zu erweitern. Wir sind zuversichtlich, dass mit dieser Studie ein erster Schritt hin zu einer einfacheren Erschließung des Inhalts keilschriftlicher Quellen genommen ist.

Fußnoten

1. Zugänglich unter http://semsearch.ukp.informatik.tu-darmstadt.de.
2. http://oracc.museum.upenn.edu
3. http://www.hethiter.net
4. Die gesamte Verarbeitungspipeline wurde hier veröffentlicht: https://github.com/UKPLab/DHd2017-semsearch-cuneiform
5. Die Listen können hier eingesehen werden: https://github.com/UKPLab/DHd2017-semsearch-cuneiform
6. Das Geschlecht wurde nicht erfasst, wir beziehen uns jeweils auf alle Teilnehmerinnen und Teilnehmer.

Bibliographie

Barthélemy, François (1998): „A morphological analyzer for akkadian verbal forms with a model of phonetic transformations“, in: *Proceedings of the Workshop on Computational Approaches to Semitic Languages* 73–81.

Cohen, Jonathan / Duncan, Donald / Snyder, Dean / Cooper, Jerrold / Kumar, Subodh / Hahn, Daniel / Chen, Yuan / Purnomo, Budirijanto / Graettinger, John (2004): „iClay: Digitizing Cuneiform“, in: *Proceedings of the International conference on Virtual Reality, Archaeology and Intelligent Cultural Heritage* 135–143.

Eckart de Castilho, Richard / Gurevych, Iryna (2014): „A broad-coverage collection of portable NLP components for building shareable analysis pipelines“, in: *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT at COLING* 1–11.

Gurevych, Iryna / Eckle-Kohler, Judith / Matuschek, Michael (2016): *Linked Lexical-Semantic Knowledge Bases: Foundations and Applications*. Morgan & Claypool Publishers.

Homburg, Timo / Chiarcos, Christian (2016): „Word Segmentation for Akkadian Cuneiform“, in: *Proceedings of the International Conference on Language Resources and Evaluation.*

Jaworski, Woyciech (2008): „Contents Modelling of Neo-Sumerian Ur III Economic Text Corpus“, in: *Proceedings of the International Conference on Computational Linguistics* 369–376.

Kataja, Laura / Koskenniemi, Kimmo (1988): „Finite-state description of semitic morphology: a case study of Ancient Akkadian“, in: *Proceedings of the Conference on Computational Linguistics* 313–315.

Lesk, Michael (1986): „Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone“, in: *Proceedings of the Annual International Conference on Systems Documentation* 24–26.

Liu, Yudong / Burkhart, Clinton / Hearne, James / Luo, Liang (2015): „Enhancing Sumerian Lemmatization by Unsupervised Named-Entity Recognition“, in: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics* 1446–1451.

Tyndall, Stephen (2012): „Toward Automatically Assembling Hittite-language Cuneiform Tablet Fragments into Larger Texts“, in: *Proceedings of ACL-2012* 243–247.

Wilhelm, Gernot (2008): „Die Edition der Keilschrifttafeln aus Boğazköy und das Projekt ‚Hethitische Forschungen‘ der Akademie der Wissenschaften und der Literatur, Mainz“, in: Wilhelm, G. (ed.): *Hattuša - Boğazköy: Das Hethiterreich im Spannungsfeld des Alten Orients*. Wiesbaden: Harrassowitz 73–86.

The Colorized Dead: Computerunterstützte Analysen der Farblichkeit von Filmen in den Digital Humanities am Beispiel von Zombiefilmen

Pause, Johannes
johannes.pause@nowalkowski.de
Technische Universität Dresden, Deutschland

Walkowski, Niels-Oliver
walkowski@nowalkowski.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland / KU Leuven, Belgien

Ein Aspekt, der zunehmende Aufmerksamkeit in der jüngeren Geschichte algorithmischer und statistischer Filmanalyse gewinnt, ist die Analyse der Farbigkeit von Filmen. Ein frühes Beispiel ist die stark rezipierte Abschlussarbeit des Grafikdesignstudenten Frederick Brodbeck (2011) (siehe Abbildung 1), der für eine Sequenz von Frames aus verschiedenen Filmen die dominanten Farben analysiert, aneinanderreiht und so ein Farbprofil des Films erzeugt. Einen ähnlichen Ansatz verfolgen Dillon Baker (2015) sowie Burghardt (2016). Die genannten Beispiele weisen sind jedoch trotz des interessanten Einblicks den sie bieten nicht unproblematisch.

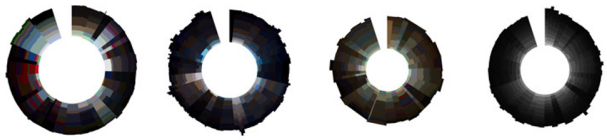


Abbildung 1: Beispiele eines Farbclusterings verschiedener Filme bei Brodbeck (2011)

Im Fall von Brodbeck und Burghard t wird für die Bestimmung der dominanten Farben in einem Frame der Clustering-Algorithmus K-Means verwendet [1] . Dieser Algorithmus ist die gängigste Strategie zur Farbquantifizierung und findet sich auch in den einschlägigen Computer Vision Bibliotheken wie zum Beispiel OpenCV [2] wieder . Die Problematik von K-Means im Kontext der filmwissenschaftlichen Farbanalyse ist vielfältig. Sie beginnt damit, dass dem Algorithmus vorgegeben werden muss, wie viele Farbcluster er erzeugen soll. Somit kann er bei der automatisierten Anwendung von bis zu 180.000 Frames pro Film nicht dem Umstand Rechnung tragen, dass es etwa farblich komplexere und einfachere Frames gibt.

Tatsächlich gibt es ein überwachtes Verfahren in dem K-Means in einer Schleife mit unterschiedlicher Clusteranzahl auf den selben Frame angewendet wird und im Sinne des sogenannten *Silhouette Koeffizienten* das bester Ergebnis bestimmt wird. Allerdings entspricht ein Clusteringergebnis bei dem die Clusterzentren bei weitestgehender Kompaktheit möglichst weit voneinander entfernt sind nicht unbedingt dem filmwissenschaftlich brauchbarsten Ergebnis (siehe Abbildung 2).

Abbildung 2: Clusteranalyse eines Frames aus The Walking Dead mit 2 bis 9 Clustern. Die Anzahl von 2 Clustern erzeugt den besten Silhouette Koeffizienten.

Intuitiv wichtige Farben eines Bildes gehen verloren. Dieses generelle Problem der Anwendung von K-Means lässt sich besonders eindrucksvoll an dem roten Mädchen aus *Schindlers Liste* aufzeigen (siehe Abbildung 3). Ein Grund dafür ist die Tatsache, dass K-Means eine Tendenz zur Bildung gleichgroßer Cluster aufweist. Folglich präferiert K-Means ein auf Verteilung hin ausgerichtetes Konzept von Dominanz.

Ein weiteres Problem von K-Means ist der Umstand, dass der Algorithmus bei jeder Anwendung leicht variierende Ergebnisse erzeugt, wobei die Variation im häufig verwendeten Spektrum zwischen 3 und 5 Clustern am größten ist. Dies kann dazu führen, dass bei einer Anwendung ein Farbton vertreten ist, der in einem weiteren Durchlauf in anderen Clustern aufgeht. Ebenfalls bleibt in den bisherigen Projektkontexten der Umstand unreflektiert, das K-Means unterschiedlich clustered, je nachdem mit welchem Farbraummodell der Frame repräsentiert wird.

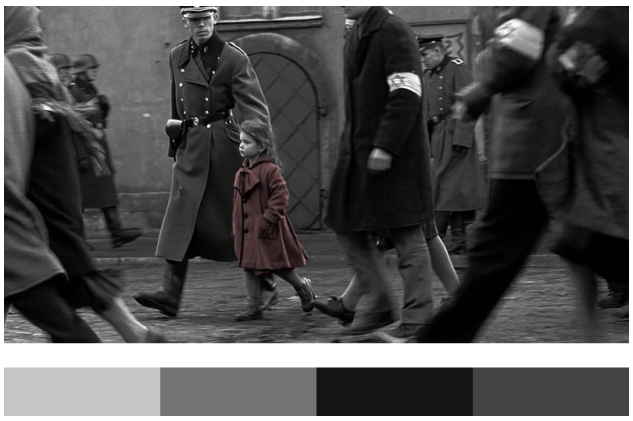


Abbildung 3: Farbclustering eines Ausschnitts aus Schindlers Liste

Das Problem der Präferenz für eine gleichmäßige Größe von Clustern sowie das zuletzt genannte zeigen die Notwendigkeit auf, die Idee einer dominanten Farbe im Kontext der computerunterstützten Filmanalyse stärker zu diskutieren. Dies ist bisher jedoch nur unzureichend erfolgt. Ein Mitgrund hierfür ist der Umstand, dass die genannten Projekte keine filmwissenschaftliche Deutung in Zusammenhang mit ihren Entwicklungen publiziert haben. Dadurch bleibt völlig offen, welche Semantik die erzeugten Muster tragen und inwieweit sie die Interpretation von Filmen beziehungsweise Filmkorpora inspirieren können. Dieser eher theoretische Problematik läßt sich auch nicht durch ein Ausweichen auf andere Clustering-Algorithmen wie DBSCAN oder Verfahren wie hierarchisches Clustering entgehen.

Angesichts der genannten Probleme erscheint die Entwicklung eines Ansatzes vonnöten, der einen technisch weniger anfechtbaren Ausgangspunkt für eine computergestützte Untersuchung von Farben im Film liefert und zugleich Anschlüsse für mögliche Interpretationen der Werke bereitstellt. Ein in diesem Zusammenhang produktives Konzept könnte das Modell der “Sieben Farbkontraste“ des Bauhaus-Künstlers und Kunstpädagogen Johannes Itten darstellen (Itten 1961: 36-109), welches die Strukturen von Farblichkeit innerhalb eines Bildes zu systematisieren erlaubt. Ausgehend von der Grundannahme, dass Farben ihre Wirkung immer in Abhängigkeit von anderen im Blickfeld befindlichen Farben entfalten, unterscheidet Itten sieben grundlegende Kontrasttypen:

- den Farbe-an-sich-Kontrast, in dem ungetrübte und daher deutlich unterscheidbare Primär-, Sekundär- oder Spektralfarben aufeinander stoßen,
- den Hell-Dunkel-Kontrast,
- den Kalt-Warm-Kontrast,
- den Qualitätskontrast, der zwischen gesättigten und trüben Farben entsteht,
- den Quantitätskontrast, der sich aus der Größe der gegenübergestellten Farbflächen ergibt,
- den Komplementärkontrast sowie
- den diesem entgegengesetzten Simultankontrast, in dem gerade das Fehlen einer Komplementärfarbe zur subjektiven Verzerrung der dargestellten Farbflächen führt.

Jeder dieser Kontrasttypen ist nach Itten mit spezifischen wirkungsästhetischen Einsatzmöglichkeiten verknüpft: So steuern sie etwa die Aufmerksamkeit der Zuschauer, ermöglichen Raumwirkungen, schaffen Orientierung oder Desorientierung, unterstützen die symbolische Semantik der Bilder oder lösen Assoziationen und Emotionen aus. Auch wenn sich die meisten dieser Effekte nicht generalisieren lassen, erscheint hier eine allgemeine rezeptionsästhetische Beschreibung doch eher möglich als bei einer Interpretation von Einzelfarben (etwa Rot als Signalfarbe, Blau als Symbol für Trauer oder Tod usw.), wie sie in der Filmanalyse bis heute Einsatz findet (etwa bei Marschall 2009).

Eine computergestützte Bestimmung nicht nur des generellen Farbclusters eines Filmes, sondern der in ihm angelegten wesentlichen Kontrasttypen kann einen ersten Ansatz für eine differenzierte Interpretation filmischer Farbschemata liefern. So kann ein Film etwa durch einen über den gesamten Filmverlauf hinweg stabilen Gegensatz von warmen und kalten Farben gekennzeichnet sein, auf der Ebene des Hell-Dunkel-Kontrastes aber eine deutlich progressive Dynamik aufweisen (zu progressiven und synopitschen Farbschemata vgl. Wulff 1988) und in wenigen besonderen Szenen starke Komplementärkontraste verwenden. Eine auf diese Weise ausbuchstabierte Entschlüsselung der komplexer Farbaspekte eines Films ließe sich dabei einerseits im Rahmen eines close readings zurate ziehen, indem etwa die dynamischen Aspekte der Farbgestaltung auf die Erzählstruktur des Werkes bezogen oder mit der Analyse inhaltlicher Leitmotive, bestimmter Figuren, dominanter Montageformen oder des Mise en Scènes verbunden werden (zur generellen Problematik der Interpretation vgl. Flückiger 2011).

Andererseits ließe sich eine computergestützte Kontrastanalyse für einen synchronen oder diachronen Vergleich mehrerer Einzelwerke oder ganzer Werkgruppen einsetzen: So ließe sich etwa überprüfen, ob sich der spezifische Stil eines Autorenfilmers auch in einem besonderen Farbprofil niederschlägt, ob sich nationale Kinematographien durch ihre Farbigkeit unterscheiden lassen oder ob sich für bestimmte Genres innerhalb konkreter Zeitspannen ein charakteristischer Einsatz besonderer Kontrastmomente nachweisen lässt.

Die Farbanalyse von Filmen in Form von Kontrasten bietet in der Umsetzung ebenfalls eine Reihe von Vorteilen gegenüber der zuvor beschriebenen Verfahrensweise. Zunächst

einmal erlauben einige Kontrastarten bereits die Untersuchung von Merkmalen der Farbsprache des Films vor der Identifikation zentraler Farben eines Frames und damit jenseits der Anwendung genannter Cluster-Algorithmen. Dies wird dadurch möglich, dass spezifische Repräsentationen des Farbraums Farben in Eigenschaften zerlegen, die direkt versuchen bestimmte Kontraste zu simulieren oder aus denen sich Kontraste leichter ableiten lassen. Am einsichtigsten ist dies im *HSV-Farbmodel* (Hue, Saturation, Value).

So is Value eine Form der Darstellung des Hell-Dunkel-Kontrast, während Saturation den Qualitätskontrast beschreibt. Zu beachten ist hier jedoch auch, dass der Value Wert nicht vollständig identisch mit der Hell-Dunkel Empfindung eines durchschnittlichen Filmbetrachters ist. Eine Übersetzung in einen solch empfunden Hell-Dunkel Kontrast ist jedoch möglich.

Die Dynamik eines bestimmten Kontrastes in einem Film kann nun erzeugt werden, indem für jeden extrahierten Frame ein Histogramm für den entsprechenden Kontrasttyp generiert wird. Die Sequenz dieser Werte lässt es zu, Muster in der Gestaltung dieses Kontrastes innerhalb des Films oder eines Filmkorpus zu identifizieren. Dabei bieten unterschiedliche Darstellungsweisen der Histogrammergebnisse in Kombination mit weiteren Phänomenen wie zum Beispiel der Berechnung des *mean absolute deviation* weiteren Deutungsspielraum. Ein Beispiel für eine sequenzielle Aneinanderreihung von Histogrammen eines Kontrasttyps als Scatterplot zeigt Abbildung 4.

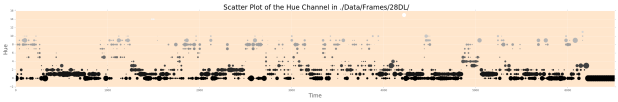


Abbildung 4: Scatterplot einer Histogramm-basierten Analyse des Hue Kontrasts in 28 Days Later

Nicht alle Kontraste kommen ohne die Bestimmung eines als absolut verstandenen Farbwertes aus. Dies ist zum Beispiel beim Simultan- und Komplementärkontrast der Fall. Der vorgestellte Ansatz ist auch nicht als Ersetzung von Clusteringverfahren zu verstehen. Beide Verfahren können auch komplementär eingesetzt werden. So kann eine Analyse des Farbe-an-sich-Kontrastes der Schwierigkeit entgegenwirken, dass K-Means einer vorgegebenen Clusterzahl folgt, bzw. der Silhouette Koeffizient keine für die Filminterpretation brauchbaren Ergebnisse produziert. Konkret kann ein hoher Farben-an-sich-Kontrast zum Anlass für die Bestimmung einer größeren Clusterzahl und umgekehrt genutzt werden. Desweiteren erlaubt das Verständnis von Farbigkeit als Kombination von Kontrasten das eingangs angesprochene Phänomen zu untersuchen und produktiv anzuwenden, dass K-Means für Daten die unterschiedlichen Farbraummodellen folgen unterschiedliche Ergebnisse liefert.

Der Hauptpunkt dieses Ansatzes ist es nicht ‘objektiv richtigere’ Clusteringergebnisse zu bekommen, sondern auf der Grundlage der Erkenntnis das es kein ‘richtiges’ Clustering von Farben in Farbkompositionen geben kann, hilfreiche und interpretierbare Ergebnisse mit gleichen und komplementären Verfahren zu produzieren. Der Kern von Ittens Herangehensweise an Farbkompositionen ist die Erkenntnis, dass ihre Analyse im wesentlichen ein wahrnehmungstheoretisches Problem ist. Entsprechend kann es erfolgreicher sein, bei der Herausarbeitung bestimmter Dimensionen von Farblichkeit zu beginnen anstatt die Varianz in diesen Dimensionen durch Clustering vor jeglicher Analyse zu reduzieren. Die angedeuteten Verfahren zeigen, dass es technisch gesehen nicht schwierig sein muss, diesem Umstand innerhalb einer Digital Humanities Perspektive Rechnung zu tragen. Der Vortrag wird die aufgezeigten Probleme bei der Farbanalyse von Filmen, den vorgestellten alternativen Ansatz sowie die Brauchbarkeit dieses Ansatzes als unterstützendes Rahmenwerk für die Filminterpretation an Hand der drei Zombiefilme 28 Days Later , [REC] und World War Z vorstellen und illustrieren.

Fußnoten

1. Baker erzeugt lediglich ein Frame-Mittelwert und benötigt daher kein Clustering
2. http://opencv.org

Bibliographie

Brodbeck, Frederic (2011): *CINEMETRICS — film data visualization*. http://cinemetrics.fredericbrodbeck.de/ [letzter Zugriff 26. August 2017].

Burghardt, Manuel / Kao, Michael / Wolff, Christian (2016): „Beyond Shot Lengths – Using Language Data and Color Information as Additional Parameters for Quantitative Movie

Analysis“, in: *DH2016: Conference Abstracts* 753–755.

Dillon Baker (2015): „Spectrum“. *Dillon Baker*. http://dillonbaker.com/spectrum/ [letzter Zugriff 16. Mai 2016].

Flückiger, Barbara (2011): „Die Vermessung ästhetischer Erscheinungen“, in: Hediger, Vinzenz / Stauff, Markus (eds.): *Zeitschrift für Medienwissenschaften* 5: 44–60.

Itten, Johannes (1961): *Kunst der Farbe*. Ravensburg: Otto Maier Verlag.

Marschall, Susanne (2009): *Farbe im Kino*. Marbug: Schüren Verlag.

Wulff, Hans J. (1988): „Die signifikativen Funktionen der Farben im Film“, in: *Kodikas/ Code* 11 (3–4): 363–376.

Von sammlungsspezifischen Visualisierungen zu nachnutzbaren Werkzeugen

Glinka, Katrin
glinka@fh-potsdam.de
Fachhochschule Potsdam

Pietsch, Christopher
cpietsch@gmail.com
Fachhochschule Potsdam

Dörk, Marian
doerk@fh-potsdam.de
Fachhochschule Potsdam

Einleitung

Die Entwicklung digitaler Werkzeuge lässt sich als wichtiger Teilbereich in den Digital Humanities identifizieren (Davis und Kräutli 2015; Schnapp et al. 2009). Entsprechende Forschung und Projektarbeit steht dabei komplexen Herausforderungen gegenüber. Nicht nur die Frage nach verfügbaren Daten, methodologischer Fundierung und technologischer Umsetzbarkeit, sondern auch die Frage nach deren langfristigen Verfügbarmachung sind wiederkehrende Themen in den Diskursen der letzten Jahre. Abgesehen von textbasierten Anwendungen in Digital Humanities (Cheema et al. 2015) etablieren sich zunehmend Projekte und Forschungsfragen im Bereich der Kunst- und Bildwissenschaften als Digital Art History (Bentkowska-Kafel et al. 2005; Drucker 2013; Promey und Stewart 1997). Eine zentrale Rolle für die Sicherstellung von Qualität und Anwendbarkeit der digitalen Werkzeuge, die sowohl für textbasierte Forschung als auch im Bereich der Kunstgeschichte und Bildwissenschaften entwickelt werden, ist die Einbindung von Forscher_innen der jeweiligen geisteswissenschaftlichen Disziplinen im Entwicklungsprozess (Drucker 2013). Gleichzeitig lässt sich die zentrale Bedeutung von Interfacedesign, Nutzungsanleitungen und Benutzerfreundlichkeit als wichtige Faktoren für die Etablierung von digitalen Werkzeugen im Forschungsprozess feststellen (Gibbs und Owens 2012). Doch selbst wenn diese Herausforderungen bewältigt werden und ein digitales Werkzeug (erfolgreich) entwickelt wurde, stellt sich weiterhin die Frage, wie die langfristige Nachnutzung im Sinne einer digitalen Nachhaltigkeit sichergestellt werden kann. Am Beispiel des Entstehungsprozesses einer sammlungsspezifischen Visualisierung und deren Weiterentwicklung zu einem nachnutzbaren Werkzeug für diverse Bildbestände werden einige zentrale Aspekte der beeinflussenden Faktoren und Lösungsansätze vorgestellt. Unser Beitrag stellt sich somit der Frage, wie sichergestellt werden kann, dass digitale Tools auch über die Laufzeit von Förderprojekten hinaus (und unabhängig von spezifischen Use-cases) dauerhaft nutzbar und weiterentwickelbar sind.

Visualisierung der Zeichnungen Friedrich Wilhelms IV.

Im Rahmen des dreijährigen BMBF Forschungsprojektes “Visualisierung kultureller Sammlungen (VIKUS)” wurde an der Fachhochschule Potsdam (FHP) in Kooperation mit der Stiftung Preußische Schlösser und Gärten (SPSG) ein webbasierter Prototyp [1]entwickelt, welcher historische Zeichnungen von Friedrich Wilhelm IV. aus den Beständen der Graphischen Sammlung der SPSG visualisiert und in einem explorativen Interface verfügbar macht (Glinka et al. 2016). Die Entwicklung der Visualisierungsumgebung zu Friedrich Wilhelm IV., abgekürzt FW4, war zunächst

spezifisch auf diesen Sammlungsbestand zugeschnitten und wurde kollaborativ in interdisziplinären Workshops vorangetrieben. Während des Entwicklungsprozesses wurden Zwischenergebnisse, Wireframes, Mock-Ups und Vorüberlegungen auf öffentlichen Workshops, im Rahmen von Vorträgen, auf Konferenzen und als Teil einer Ausstellung präsentiert und diskutiert. Bereits zu diesem Zeitpunkt zeichnete sich der Bedarf und das Interesse anderer Sammlungsinstitutionen und Forschungsgruppen an den Funktionen und Darstellungsmodi der Visualisierung ab. Trotz der ursprünglichen Entwicklungsfokussierung auf den FW4-Sammlungsbestand, stellten sich grundlegende Funktionen und Strukturen als potenziell generalisierbar und auf andere Bestände übertragbar heraus. Zentral hierbei ist die generische Struktur, entlang derer nicht nur die Sammlung von Friedrich Wilhelms Zeichnungen organisiert ist, sondern welche ein zentrales Ordnungselement zahlreicher Sammlungen darstellt: zeitliche Einordnung und Verschlagwortung. Somit entwickelte sich aus dem Ansatz einer sammlungsspezifischen Visualisierungsumgebung das Konzept für ein digitales Werkzeug. Diese nächste Entwicklungsstufe erprobte zunächst die Übertragung auf vier weitere Bestände in enger Kooperation mit den jeweiligen Sammlungsinstitutionen und folgt dabei dem Anspruch, die technologischen Lösungen und darstellerischen Optionen langfristig und nachhaltig für zahlreiche weitere Anwendungsfälle zugänglich zu machen. Die entwickelte Visualisierungsumgebung wird als *VIKUS Viewer* (kurz: »VV«) anderen Sammlungsinstitutionen zur dauerhaften Nutzung zur Verfügung stehen.

Von FW4 zu »VV«

Die Grundstruktur der Visualisierungsumgebung zu den Zeichnungen Friedrich Wilhelms IV. beruht auf einer zeitlichen Anordnung und der Verknüpfung mit Stichworten. Im Browserfenster werden die 1492 Zeichnungen auf einem dynamischen Canvas entsprechend ihres Entstehungsjahrs auf einer Zeitleiste arrangiert. Die Darstellung ähnelt dabei einem klassischen Balkendiagramm, wobei jedoch jeder Balken aus tatsächlichen Digitalisaten einzelner Zeichnungen zusammengesetzt ist. Hierdurch ist auf den ersten Blick erkennbar, welches zeitliche Spektrum die Sammlung abdeckt und welche Menge an Objekten diesem Zeitraum zugeordnet sind. Die Schlagworte wiederum sind am oberen Rand angeordnet, alphabetisch sortiert und geben über die Schriftgröße darüber Auskunft, wie häufig sie Objekten zugeordnet worden sind. Gleichzeitig fungieren die Schlagworte als Filter. Bei Auswahl eines oder mehrerer Schlagworte werden die mit dem Begriff verknüpften Objekte oberhalb der Zeitleiste angezeigt, alle anderen unterhalb. Somit lassen sich themenbezogene Häufungen oder das Aufkommen von thematischen Fokussierungen in ihrem zeitlichen Kontext ablesen. Das gesamte Arrangement ist zugleich ein stufenlos zoombares Interface. In jeder der beschriebenen Anordnungen kann also von der Übersicht (“distant viewing”) in einer kontinuierlichen Bewegung in die hochauflösende Ansicht des Digitalisats (“close viewing”) gezoomt werden. In dieser Detailansicht werden automatisch in einem Textpanel die dem Objekt zugeordneten Metadaten und beschreibende Texte eingeblendet [2].

Für die Weiterentwicklung dieser sammlungsspezifischen Visualisierung FW4 zum nachnutzbaren Werkzeug, dem *VIKUS Viewer* »VV«, ergeben sich schließlich eine Reihe an Fragestellungen und Herausforderungen.

Um den »VV« für andere Sammlungen nutzbar zu machen, muss dieser von der spezifischen Metadatenstruktur des FW4-Bestandes losgelöst werden. Trotz des Anspruchs, den VIKUS Viewer für eine Vielzahl an Beständen nutzbar zu machen, kann nicht jeder Sonderfall mit objektspezifischen Herausforderungen und Sammlungsgrößen in Betracht gezogen werden. Um die Potenziale und Limitationen des »VV« zu erproben, wird die Generalisierung daher zunächst auf Basis von weiteren Beständen der SPSG und anderen Projektpartnern vorangetrieben. Hierbei ergeben sich als Minimalanforderungen an die vom »VV« abgedeckten Sammlungen eine Liste an Eigenschaften. Dazu zählen primär die zeitliche Einordnung und die Verschlagwortung entlang eines kontrollierten Vokabulars. Zusätzlich sollte der Bestand in ausreichend guter Qualität digitalisiert sein (und als jpg vorliegen), um die Zoomfunktion (von der Übersicht ins Detail) voll ausnutzen zu können. Die benötigten Sammlungsdaten müssen in einem standardisierten CSV Format vorliegen, welches ein gängiges und einfach zu erstellendes Datenformat ist. Auch wenn dies bereits eine recht breite Nutzung erlaubt, werden darüber hinaus Lösungen für weitere Erweiterungen und Ansprüche entwickelt, welche in Kombination die langfristige und nachhaltige Nutzung und

Verfügbarmachung des »VV« gewährleisten sollen.

Skalierbarkeit: Während der Bestand zu FW4 nur 1492 Zeichnungen umfasst, soll der »VV« für Sammlungen optimiert werden, die bis zu 7000 Objekte umfassen. Weiterhin sollen Konzepte auf Machbarkeit überprüft werden, welche sogar Darstellungslösungen für größere Bestände bieten. Hierzu zählen aggregierte Übersichten, in denen die eigentlichen Digitalisate erst nach einer Themeneinschränkung angezeigt werden. Bei der Skalierbarkeit orientiert sich die Entwicklung des VV an den letzten beiden Generationen von Laptops mittlerer Leistungsstärke, um eine breite Einsatzfähigkeit zu sichern.

Alternative Ansichten und Gestaltung: zusätzliche Visualisierungsformen wie z.B. Anordnungen auf Basis von Metadatenähnlichkeit erweitern zudem die abbildbaren Facetten der verschiedenen Sammlungen. Ebenso lassen sich im »VV« gestalterische Elemente wie Hintergrundfarbe, Textfarbe und -größe, Schriftart, Linkfarben etc. anpassen, so dass auf die visuellen Eigenschaften der dargestellten Sammlungen eingegangen werden kann.

Gezielte Suche: Der Fokus bei FW4 lag auf der Visualisierung, die sich in diesem Fall als Ergänzung zum bereits bestehenden eher klassischen digitalen Bestandskatalog versteht. Da im Bestandskatalog als primärer Zugang u.a. eine Suchfunktion angeboten wird, wurde diese Funktion nicht als Teil der FW4 Visualisierungsumgebung implementiert. Um den VIKUS Viewer von ergänzenden Zugängen unabhängig zu machen, wird eine gezielte Suchfunktion eingebunden. Die Suchanfragen werden in einer Rückkopplung auf die Darstellung und die Anordnung des Bestandes Einfluss nehmen und somit visuell nachvollziehbar sein.

Mehrseitigkeit: Über die Darstellung von zweidimensionalen und einseitigen Objekten (wie Zeichnungen oder Gemälde) hinaus, soll der »VV« sowohl dreidimensionale Objekte, welche beispielsweise in mehransichtigen Abbildungen digitalisiert wurden, als auch mehrseitige Schriften darstellen können. In der Detailansicht können über verschiedene Darstellungsmodi die weiteren Abbildungen zum Objekt bzw. die zusätzlichen Seiten im Detail betrachtet werden.

Implementierung: Die FW4-Visualisierung basiert auf einem komplexen Einsatz von innovativen Webtechnologien, deren Kenntnis nicht als Voraussetzung für die Anwendung auf neue Bestände vorausgesetzt werden können. Somit wird für den VIKUS Viewer ein zugänglicher Workflow entwickelt, der in Kombination mit detaillierten Anleitungen die Implementierung erleichtert und für weniger technisch versierte Nutzergruppen öffnet. Hierzu zählt eine ausführliche Dokumentation und Anleitung auf GitHub, wie der »VV« implementiert werden kann.

Langfristige Verfügbarmachung: Die prototypische Umsetzung der FW4-Visualisierung ist bereits mit offenen und langlebigen Web-Standards (HTML, CSS und JavaScript) erfolgt. Ebenso ist es das Ziel, dass Bestände, die mittels »VV« in Zukunft verfügbar gemacht werden, langfristig abrufbar und verlässlich archiviert werden können. Die Komponenten werden in der JavaScript library *react* geschrieben, durch dieses modulbasierte Programmieren und die Einhaltung von gängigen Standards wird die Erweiterbarkeit sichergestellt. Die hauptsächlich verwendeten JavaScript libraries wie d3.js, pixi.js, react, node, u.a. sind in der Open Source Community weit verbreitet. Auch wenn dies keine langfristige Nachnutzbarkeit *garantiert*, werden die libraries jedoch von einer großen Community kontinuierlich weiterentwickelt und bieten somit eine relativ hohe Verlässlichkeit in Bezug auf langfristige Funktionalität.

Open Source: Neben der Publikation des Quellcodes im offenen online-code-repository GitHub wird der VIKUS Viewer von einer ausführlichen Dokumentation begleitet, welche die Weiterentwicklung und Nutzung des Codes ermöglicht und nachvollziehbar macht. Dies wird unterstützt durch aktives Community-Building wie z.B. Workshops und Einführungen in die Konfiguration und Nutzung des »VV«.

Diskussion

Dieser letzte Aspekt erweitert die vorangegangenen Überlegungen zu den Funktionalitäten des VIKUS Viewers auf allgemeinere Aspekte, welche im Zuge des Entwicklungsprozesses relevant wurden und weiterhin verhandelt werden. Selbst wenn ein Werkzeug als Open Source Projekt entwickelt wird, hängt das langfristige Überleben von einer aktiven Community und entsprechender Ressourcen ab. Die Vernetzung und der offene Austausch innerhalb einer Forschungsgemeinde sowie die enge Zusammenarbeit mit Institutionen, die ebenso einen praktischen Nutzen von einem solchen Werkzeug haben, erleichtert über die Bekanntheit eines Tools dessen kontinuierliche Weiterentwicklung und Nutzung. Dies setzt das offene Teilen von

(Zwischen-)Ergebnissen auch während des Entwicklungsprozesses voraus. Gleichzeitig ist auf Seiten der Förderinstitutionen zu beachten, dass sie ebenso eine zentrale Rolle in Bezug auf digitale Nachhaltigkeit einnehmen. So kann beispielsweise durch zielgerichtete Anschlussförderung von vielversprechenden Ergebnissen, welche in der Forschungsgemeinschaft und bei Sammlungsinstitionen auf eine gewisse “Nachfrage” stoßen, eine stabile, nachhaltige und zielgruppenorientierte Implementierung und Publikationen von Tools unterstützt und sichergestellt werden. Ebenso ist es erstrebenswert über die Einbindung verschiedener institutioneller und wissenschaftlicher Akteure, welche Interesse an der Weiterentwicklung und Anpassung von digitalen Werkzeugen haben und somit zu längerfristigen Kooperationspartnern werden, die Nachnutzung sicherzustellen. Somit soll der Weg dafür geebnet werden, dass das entwickelte digitale Angebot auch nach Ablauf der Projektförderung und somit möglicherweise der Veränderung von Team- und Mitarbeiterstrukturen aufrechterhalten werden kann. Bei der Entwicklung experimenteller DH-Tools haben wir bereits mehrfach gute Erfahrungen mit heterogenen und dynamischen Teams gemacht, die sich neben wissenschaftlichen Mitarbeiter_innen ebenso aus Studierenden und Freelancern zusammensetzen. Dabei stellt die Finanzierung solcher freier Forscher_innen noch die Seltenheit dar und stößt regelmäßig an Grenzen, wenn es um die Akquise der entsprechenden Mittel geht. Der kürzlich von der Open Knowledge Foundation Deutschland und dem Bundesministerium für Bildung und Forschung ausgeschriebene Prototype Fund [3]ist eine erfreuliche Ausnahme und zeigt in die richtige Richtung.

[1] https://uclab.fh-potsdam.de/fw4/

[2] Für eine detailliertere Beschreibung der Funktionalität und des Interaktionsmodells siehe Glinka, Katrin / Pietsch, Christopher / Dilba, Carsten / Dörk, Marian (2016): “Linking structure, texture and context in a visualization of historical drawings by Frederick William IV (1795- 1861)”, in: *International Journal of Digital Art History*, 2.

[3] https://prototypefund.de/

Bibliographie

Bentkowska-Kafel, Anna / Cashen, Trish / Gardiner, Hazel (eds.) (2005): „Digital art history: a subject in transition“, in: *Computers and the history of art series* 1: 1. Intellect.

Boyd Davis, Stephen / Kräutli, Florian (2015): „The Idea and Image of Historical Time: Interactions between Design and Digital Humanities“, in: *Visible Language* 49 (3): 101.

Cheema, Muhammad F. / Jänicke, Stefan / Franzini, Greta / Scheuermann, Gerik (2015): „On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges“, in: *Eurographics Conference on Visualization (EuroVis) - STARs* 83–103.

Drucker, Johanna (2013): „Is There a ‚Digital‘ Art History?“, in: *Visual Resources* 29 (1–2): 5–13.

Gibbs, Fred / Owens, Trevor (2012): „Building better digital humanities tools: Toward broader audiences and user-centered designs“, in: *DHQ: Digital Humanities Quarterly* 6, 2. http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html [letzter Zugriff 1. Dezember 2016].

Glinka, Katrin / Pietsch, Christopher / Dilba, Carsten / Dörk, Marian (2016): „Linking structure, texture and context in a visualization of historical drawings by Frederick William IV (1795- 1861)“, in: *International Journal of Digital Art History* 2.

Promey, Sally M / Stewart, Miriam (1997): „Digital art history: A new field for collaboration“, in: *American Art* 11 (2): 36–41.

Schnapp, Jeffrey / Presner, Todd / Lunenfeld, Peter et al. (2009): *The digital humanities manifesto 2.0*. http://manifesto.humanities.ucla.edu/2009/05/29/the-digital-humanities-manifesto-20/ [letzter Zugriff 1. Dezember 2016].

Wiederholende Forschung in den digitalen Geisteswissenschaften

Schöch, Christof
christof.schoech@uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Die Reproduzierbarkeit von Forschungsarbeiten ist in zahleichen Disziplinen ein drängendes und vieldiskutiertes Problem. Laut einer *Nature*-Umfrage nehmen 52% der

befragten ForscherInnen eine “significant reproducibility crisis” wahr (Baker 2016). Metastudien aus der Psychologie (Bohannon 2015) oder den Wirtschaftswissenschaften (Camerer 2016) berichten von niedrigen Reproduzierbarkeitsquoten. Forderungen nach reproduzierbarer Forschung werden nicht nur in der Informatik (Mesirov 2010, Peng 2010) formuliert. Insbesondere für die empirisch und ggfs. quantitativ arbeitenden Teile der digitalen Geisteswissenschaften sind diese Debatten relevant (Padilla und Higgins 2016).

Hier stehen allerdings nicht die Anforderungen an wiederholbare Forschung im Fokus, sondern umgekehrt die Herausforderungen, vor denen wiederholende Forschung steht. Letztere ist in den digitalen Geisteswissenschaften in besonderem Maße aufschlussreich, stellt doch der Paradigmenwechsel von dominant hermeneutischen zu dominant empirischen Methoden in den Geisteswissenschaften die Kontinuität des wissenschaftlichen Diskurses auf eine Zerreißprobe. Die digitalen Geisteswissenschaft sind gefordert, die eigene Anschlussfähigkeit an etablierte Konzepte, Fragestellungen und Erkenntnisziele sicherzustellen. Studien, die vorhandene Arbeiten mit digitalen Mitteln wiederholen, platzieren diese Kontinuitätsfrage gewissermaßen unter einem Mikroskop. Zudem treten im praktischen Nachvollzug einer Originalstudie die (teils impliziten) Annahmen sowie die Stärken und Grenzen beider Ansätze plastisch hervor. So versprechen Wiederholungsstudien inhaltlichen ebenso wie methodischen Erkenntnisgewinn (vgl. Rockwell 2016).

Auf eine konzeptuellen und begrifflichen Klärung zum beschrieben Problemfeld der wiederholenden Forschung folgen im hier skizzierten Beitrag zwei unterschiedliche literaturwissenschaftliche Fallstudien, in denen vorhandene Forschungsbeiträge mit digitalen Daten und Methoden wiederholt worden sind.

Typen wiederholender Forschung

Für die vielfältigen Beziehungen zwischen einer bereits vorliegenden Studie und einer diese wiederholenden Studie sind in der Forschungsliteratur zahlreiche Begriffe vorgeschlagen worden, darunter insbesondere Replikation, Reproduktion und Reanalyse (Drummond 2009, Gomez und Juristo 2010).

Zur konzeptuellen Klärung werden hier drei wesentlicheAspekte berücksichtigt: die Fragestellung, die Daten und die Analysemethode. Wiederholungsstudien unterscheiden sich, je nachdem ob Fragestellung, Daten und Methoden gegenüber der Originalstudie identisch oder verändert sind (Abbildung 1).

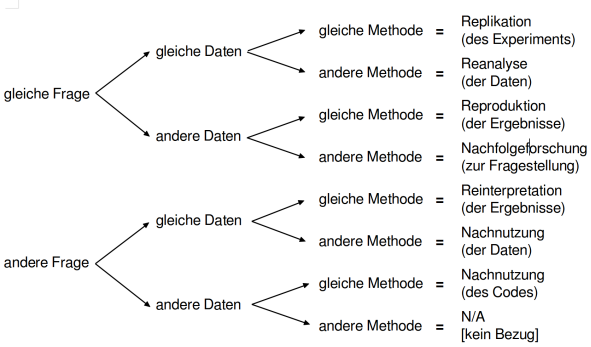


Abbildung 1: Das konzeptuelle und begriffliche Feld der wiederholenden Forschung.

Der Begriff "Replikation" bezeichnet hier die exakte Wiederholung einer Studie. Die gleiche Forschungsfrage wird mit gleicher Datengrundlage und gleichen Methoden erneut bearbeitet. Ziel ist es zu prüfen, ob die gleichen Ergebnisse ermittelt werden können, was ein Hinweis auf die korrekte Durchführung der Originalstudie ist.

Der Begriff "Reproduktion" bezeichnet eine freiere Wiederholung. Die gleiche Fragestellung wird mit den gleichen Analysemethoden, aber neu erhobenen oder erweiterten Daten durchgeführt. Ziel ist es zu prüfen, ob die Analysemethode auch mit veränderten Daten die gleichen Schlussfolgerungen erlaubt, d.h. ob die Ergebnisse generalisierbar sind.

Der Begriff "Reanalyse" bezeichnet ebenfalls eine freiere Wiederholung. Hier wird die gleiche Fragestellung mit den gleichen Daten, aber einer anderen (bspw. verbesserten oder neu implementierten) Analysemethode bearbeitet. Wird die gleiche Fragestellung sowohl mit anderen Daten als auch mit anderen Methoden bearbeitet, kann man von “Nachfolgeforschung” sprechen.

Auch wenn eine veränderte Fragestellung im Fokus steht, kann ein Bezug zu einer früheren Studie bestehen. Die Bearbeitung einer veränderten Fragestellung mit den gleichen Daten und der gleichen Methode kann als “Reinterpretation” der Ergebnisse aus einer anderen Perspektive verstanden werden. Der erneute Einsatz von Daten oder Code aus einer

früheren Studie für die Bearbeitung einer neuen Fragestellung ist eine “Nachnutzung”. Kein (hier wesentlicher) Bezug besteht, wenn Fragestellung, Daten und Code gegenüber einer früheren Studie verändert wurden.

Die folgenden beiden Fallstudien beziehen sich auf sehr unterschiedliche Originalstudien, illustrieren die spezifischen Herausforderungen, die jeweils hiermit zusammenhängen und werfen ein Schlaglicht auf das Verhältnis der digitalen Geisteswissenschaften zu früherer Forschung.

Erste Fallstudie: Richeaudeau zur Satzlänge bei Georges Simenon

Die erste Fallstudie bezieht sich auf die Wiederholung einer Studie von François Richeaudeau zur Satzlänge im umfangreichen Werk des belgischen Autors Georges Simenon. Die 1982 veröffentlichte Studie ist quantitativ angelegt, wurde allerdings nicht computergestützt durchgeführt. Zentrale These ist, dass Simenons Romanwerk sich durch die Verwendung besonders kurzer Sätze auszeichne. Dies wird als ein Faktor unter anderen interpretiert, der zum weltweiten Erfolg des Autors beigetragen hat (Richeaudeau 1982).

Obwohl in diesem Fall die Textsammlung bekannt und das verwendete Verfahren quantitativ ist, kann nur in Ansätzen eine Replikation der Studie (im oben definierten Sinne) vorgenommen werden. Beispielsweise ist nicht dokumentiert, wie Satz und Wort für die Messung der Satzlänge definiert sind. Dies musste neu entschieden und implementiert werden. Die erneute Messung der Satzlängen in den 25 von Richeaudeau untersuchten Texten Simenons anhand einer einfachen, aber angemessen erscheinenden Definition von Satz und Wort ergibt um durchschnittlich 15% niedrigere Werte (siehe Abbildung 2; Details in Schöch 2016). Das scheint zunächst Richeaudeaus These sogar noch zu stärken.

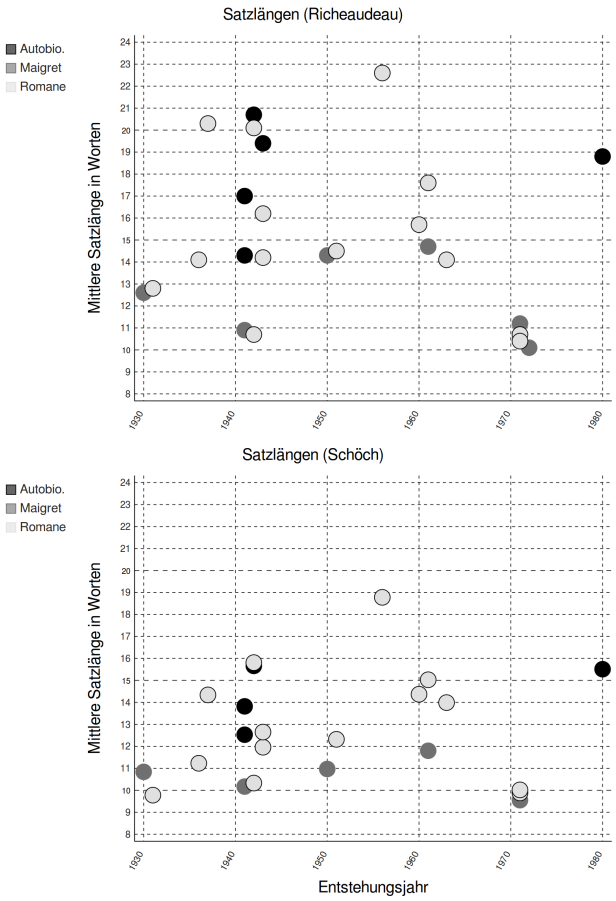


Abbildung 2: Von Richeaudeau (oben) und in der Wiederholungsstudie (unten) erhobene Satzlängen unter Verwendung der gleichen Texte.

Allerdings wird deutlich, dass 25 Werken nicht ausreichen, um Richeaudeaus weiterführende Thesen einer Entwicklung Simenons’ Stils über die Zeit (hin zu zunehmend kürzeren Sätzen in den Romanen) sowie in Abhängigkeit der von ihm praktizierten Gattungen (längere Sätze in den autobiographischen Schriften als in den Romanen) zu prüfen. Erst mit deutlich mehr Werken (hier 127 Texte) und mit Hilfe eines statistischen Signifikanztests, kann die erste dieser Thesen geprüft und widerlegt werden, die zweite dieser Thesen dagegen klar bestätigt werden (Abbildung 3).

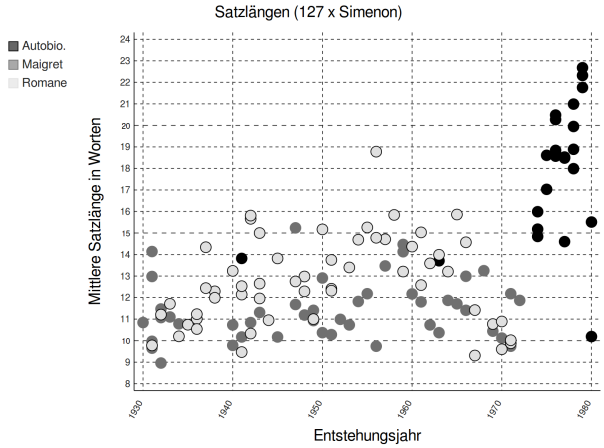


Abbildung 3: Satzlängen für 127 Werke Simenons in drei Gattungen: autobiographische Werke (schwarz), Maigret-Romane (grau), psychologische Romane (weiß). Statistisch massiv signifikanter Unterschied zwischen Romanen und autobiographischen Werken.

Zudem verfügt Richeaudeau als Vergleichsmaßstab nur über Zahlen aus einer Einzelstudie zu Marcel Proust, im Vergleich zu dessen langen Sätzen Simenons Sätze kurz erscheinen müssen. Der Vergleich mit 195 französischen Romanen, die wie Simenons Werke zwischen 1930 und 1980 erschienen sind, zeigt hingegen, dass es zwar einige wenige Romanciers gibt, die deutlich längere Sätze verwenden als Simenon, dieser aber keinesfalls ungewöhnlich kurze Sätze verwendet (Abbildung 4).

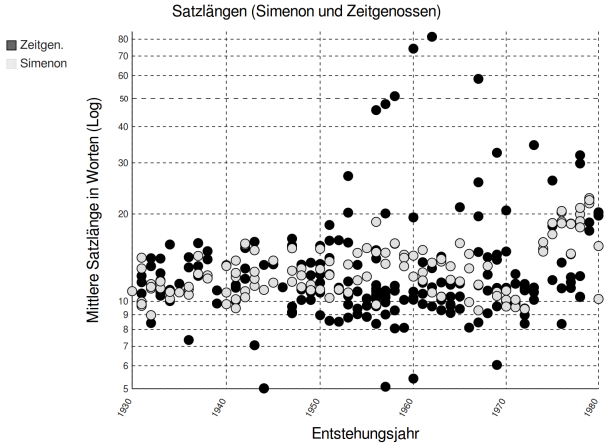


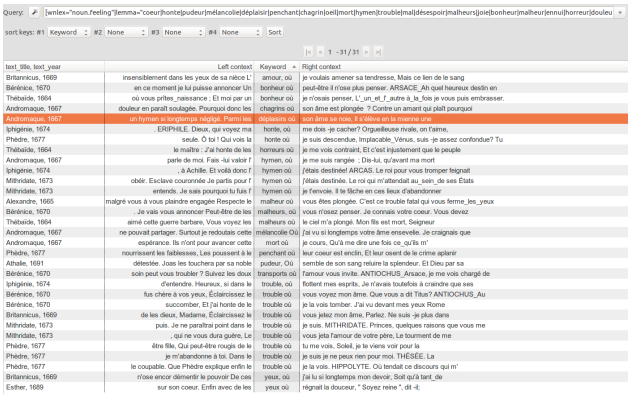
Abbildung 4: Satzlänge bei Georges Simenon (weiß) und in 195 zeitgenössischen Romanen (schwarz). Kein statistisch signifikanter Unterschied.

Abschließend kann festgehalten werden, dass hier weniger eine methodische Kluft überwunden werden musste, als vielmehr mangelnde Dokumentation des eingesetzten Verfahrens eine Herausforderung darstellt. Anders in der folgenden Fallstudie.

Zweite Fallstudie: Spitzers Stilanalyse des Werks Jean Racines

Die zweite Fallstudie bezieht sich auf die Wiederholung einer bis heute viel beachteten Stilanalyse, die der Romanist Leo Spitzer 1928 über den französischen Dramatiker Jean Racine vorgelegt hat. Spitzer verfolgt die These, dass in Racines Tragödien ein stilistischer "Dämpfungseffekt" (als Autorenstil) aufgezeigt werden kann. Offen lässt Spitzer, inwiefern dieser Effekt zugleich paradigmatisch für die Klassik (als Epochenstil) ist. Spitzer unterscheidet rund 50 stilistische Phänomene, die zum “Nüchtern-Gedämpften, Verstandesmäßig-Kühlen, fast Formelhaften” in Racines Stils beitragen. Er beschreibt sie nuancenreich und illustriert sie mit zahlreichen Beispielen. Zur Veranschaulichung seien hier nur einige Definitionen Spitzers zitiert: “die Personifizierung von Abstrakta”, “konturverwischende Plurale” oder “das entgrenzende *où*” (Spitzer 1928).

Für die Reproduktionsstudie stehen die gleichen Texte zur Verfügung, die auch Spitzer verwendet hat, allerdings in digitaler Form und anderen Textausgaben folgend. Spitzers stilistische Phänomene wurden in Form komplexer Suchabfragen, die mit Hilfe der “Corpus Query Processing”-Sprache CQP (http://cwb.sourceforge.net/files/CQP_Tutorial/) formuliert wurden, im Textanalyse-Tool TXM (http://www.textometrie.fr) nachmodelliert und quantifiziert (siehe Abbildung 5). Auch mit Hilfe aufwändiger Annotationen der Texte (morpho-syntaktische sowie semantische Annotation mit WordNet) gelang dies mit zufriedenstellender Genauigkeit nur für 30 der rund 50 von Spitzer analysierten stilistischen Phänomene.



text_title, text_year	Left context	Keyword	Right context
Britannicus, 1669	insensiblement dans les yeux de sa nièce L'	amour, où	je voulais amener sa tendresse, Mais ce lien de le sang
Bérénice, 1670	en ce moment je lui puisse annoncer Un	bonheur où	peut-être il n'ose plus penser. ARSACE_Ah quel heureux dessin en
Thébaïde, 1664	où vous prîtes_naissance ; Et moi par un	bonheur où	je n'osais penser, L'_un_et_f_autre à_la_fois je vous puis embrasser.
Andromaque, 1667	douleur en paraît soulagée. Pourquoi donc les	chagrins où	son âme est plongée ? Contre un amant qui plaît pourquoi
Andromaque, 1667	un hymen si longtemps négligé. Parmi les	déplaisirs où	son âme se noie, Il s'élève en la menne une
Iphigénie, 1674	ERIPHILE. Dieux, qui voyez ma	honte, où	me dois -je cacher? Orgueilleuse rivale, on l'aime,
Phèdre, 1677	seule. Ô toi ! Qui vois la	honte où	je suis descendue, Implacable_Vénus, suis -je assez confondue? Tu
Thébaïde, 1664	le maître : J'ai honte de les	horreurs où	je me vois contraint, Et c'est injustement que le peuple
Andromaque, 1667	parle de moi. Fais -lui valoir l'	hymen, où	je me suis rangée ; Dis-lui, qu'avant ma mort
Iphigénie, 1674	, à Achille. Et voilà donc l'	hymen où	j'étais destinée! ARCAS. Le roi pour vous tromper feignait
Mithridate, 1673	obéir. Esclave couronnée Je partis pour l'	hymen où	j'étais destinée. Le roi qui m'attendait au_sein_de ses États
Mithridate, 1673	entends. Je sais pourquoi tu fuis l'	hymen où	je l'envoie. Il te fâche en ces lieux d'abandonner
Alexandre, 1665	malgré vous à vous plaindre engagée Respecte le	malheur où	vous êtes plongée. C'est ce trouble fatal qui vous ferme_les_yeux
Bérénice, 1670	Je vais vous annoncer Peut-être de les	malheurs, où	vous n'osez penser. Je connais votre coeur. Vous devez
Thébaïde, 1664	aimé cette guerre barbare, Vous voyez les	malheurs où	le ciel m'a plongé. Mon fils est mort, Seigneur
Andromaque, 1667	ne pouvait partager. Surtout je redoutais cette	mélancolie Où	j'ai vu si longtemps votre âme ensevelie. Je craignais que
Andromaque, 1667	espérance. Ils n'ont pour avancer cette	mort où	je cours, Qu'à me dire une fois ce_qu'ils m'
Phèdre, 1677	nourrissent les faiblesses, Les poussent à le	penchant où	leur coeur est enclin, Et leur osent de le crime aplanir
Athalie, 1691	détestée. Joas les touchera par sa noble	pudeur, Où	semble de son sang reluire la splendeur. Et Dieu par sa
Bérénice, 1670	soin peut vous troubler ? Suivez les doux	transports où	l'amour vous invite. ANTIOCHUS_Arsace, je me vois chargé de
Iphigénie, 1674	d'entendre. Heureux, si dans le	trouble, où	flottent mes esprits, Je n'avais toutefois à craindre que ses
Bérénice, 1670	fus chère à vos yeux, Éclaircissez le	trouble où	vous voyez mon âme. Que vous a dit Titus? ANTIOCHUS_Au
Bérénice, 1670	succomber, Et j'ai honte de le	trouble où	je la vois tomber. J'ai vu devant mes yeux Rome
Britannicus, 1669	de les dieux, Madame, Éclaircissez le	trouble où	vous jetez mon âme. Parlez. Ne suis -je plus dans
Mithridate, 1673	puis. Je ne paraîtrai point dans le	trouble où	je suis. MITHRIDATE. Princes, quelques raisons que vous me
Mithridate, 1673	, qui ne vous dura guère, Le	trouble où	vous jeta l'amour de votre père, Le tourment de me
Phèdre, 1677	être fille, Qui peut-être rougis de le	trouble où	tu me vois, Soleil, je te viens voir pour la
Phèdre, 1677	je m'abandonne à toi. Dans le	trouble où	je suis je ne peux rien pour moi. THESÉE. La
Phèdre, 1677	le coupable. Que Phèdre explique enfin le	trouble où	je la vois. HIPPOLYTE. Où tendait ce discours qui m'
Britannicus, 1669	n'ose encor démentir le pouvoir De ces	yeux, où	j'ai lu si longtemps mon devoir, Soit qu'à tant_de
Esther, 1689	sur son coeur. Enfin avec de les	yeux où	régnait la douceur, " Soyez reine ", dit -il;

Abbildung 5: Keyword-in-Context Ansicht der Suchergebnisse einer CQP-Abfrage in TXM für das “entgrenzende où”.

Der über Spitzers Studie hinausgehende Vergleich der Häufigkeiten der Phänomene bei Racine einerseits, in einem Vergleichskorpus zeitgenössischer französischer Verstragödien andererseits, zeigt, dass überhaupt nur drei der 30 Phänomene bei Racine in statistisch signifikanter Weise überrepräsentiert sind (Abbildung 6). Trotz Spitzers Fokus auf Racine handelt es sich bei den von ihm identifizierten “gedämpften Stil” also gerade nicht um einen für Racine distinktiven Autorenstil, sondern um einen weit verbreitete Epochenstil.

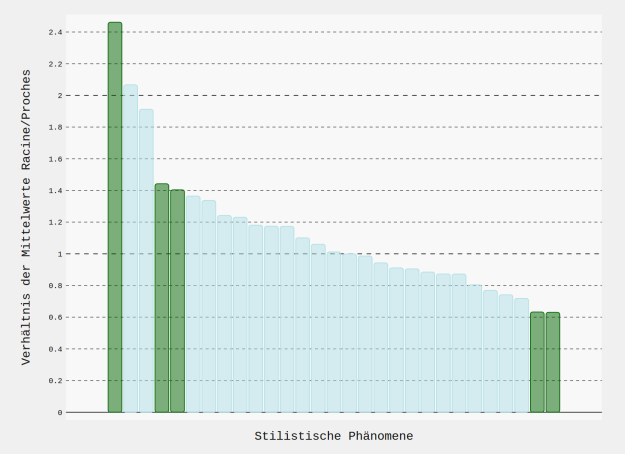


Abbildung 6: Die Häufigkeiten 30 stilistischer Phänomene im Vergleich der Tragödien Racines und 49 zeitgenössischer Tragödien. Werte > 1: bei Racine überrepräsentiert; Werte < 1: bei Racine unterrepräsentiert. Statistisch signifikant abweichende Phänomene sind hervorgehoben.

Bei der Racine-Studie liegt die wesentliche Herausforderung in der algorithmischen Modellierung stilistischer Phänomene, für deren Definition Spitzer subtile semantische Unterscheidungen und kontextuelle Informationen einsetzt, wie sie der algorithmischen Analyse derzeit nur unvollständig zugänglich sind.

Schlussfolgerungen

Beiden Fallstudien zeigen, dass Forschungsarbeiten, die verwendete Daten und Code nicht publizieren, sich nicht für eine Replikation, Reproduktion oder Reanalyse im oben definierten, engen Sinne eignen. Zu Vieles bleibt implizit, wenn das methodische Vorgehen nicht als Solches detailliert dokumentiert wurde. Das gilt auch dann, wenn Fragestellung und Methode prinzipiell einer datengestützten Wiederholung entgegenkommen (wie bei der Satzlängen-Studie).

Zugleich zeigt sich, dass stärker von der Originalstudie abweichende Nachfolgestudien es erst erlauben, eine umfangreichere Datengrundlage zu verwenden und/oder verbesserte Analysemethoden einzusetzen, wodurch sich die Aussagekraft der Analysen gegenüber der Originalstudie deutlich erhöht. Solche Studien sind zudem hilfreich, um die Anschlussfähigkeit aktueller empirischer und ggfs. quantitativer Methoden in den digitalen Geisteswissenschaften an frühere Forschung zu erproben. Und erst das bewußte, kontrollierte Abweichen von der Originalstudie macht wesentliche Grundannahmen und Erkenntnisinteressen sowohl der Originalstudie als auch der Wiederholungsstudie bewußt, beispielsweise die jeweils unterschiedlichen Stilbegriffe.

Die beiden Fallstudien hinterfragen auch die oben eingeführte binäre Opposition zwischen “identischen” und “veränderten” Fragestellungen, Daten und Methoden im Kontext solcher Wiederholungsstudien. Dennschon der Wechsel von gedruckten Texten zu digitalen Textdaten, selbst bei identischer Korpuszusammenstellung, führt zwar zu vergleichbaren, keinesfalls aber identischen Daten und verlangt auch veränderte Methoden.

Schließlich liegt nahe, dass gerade wiederholende Studien auch selbst dem Anspruch an Reproduzierbarkeit gerecht werden sollten. In diesem Sinne sind zugrundliegende Texte, Metadaten, Code und (teils interaktive) Abbildungen der hier dargestellten Wiederholungsstudien verfügbar, soweit es urheberrechtliche Einschänkungen möglich machen. Siehe https://github.com/cligs/projects (Ordner “2016/simenon” und “2016/racine”), DOI: http://doi.org/10.5281/zenodo.163223 .

Förderhinweis

Die vorliegende Arbeit wurde im Rahmen der Nachwuchsgruppe "Computergestützte literarische Gattungsstilistik" (CLiGS) erstellt, die vom BMBF gefördert wird (FKZ 01UG1508).

Bibliographie

Baker, Monya (2016): „Is there a reproducibility crisis?“, in: *Nature* 533: 452–454.

Bohannon, John (2015): „Many psychology papers fail replication test“, in: *Science Magazine* 349.6251: 910–911.

Camerer, Colin F. et al. (2016): „Evaluating replicability of laboratory experiments in economics“, in: *Science Magazine* 351.6280: 1433–1436.

Drummond, Chris (2009): „Replicability is not Reproducibility: Nor is it Good Science“, in: *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*.

Gomez, Omar S. / Juristo, Natalia / Vegas, Sira (2010): „Replication, Reproduction and Re-analysis: Three ways for verifying experimental findings“, in: *RESER ’2010*.

Padilla, Thomas / Higgins, Devin (2016): „Data Praxis in the Digital Humanities: Use, Production, Access“, in: *DH2016: Conference Abstracts* 644–646 http://dh2016.adho.org/abstracts/150 .

Peng, Roger D. (2011): „Reproducible Research in Computational Science“, in: *Science Magazine* 334: 1226–1227.

Richeaudeau, François (1982): „Simenon: une écriture pas si simple qu'on le penserait“, in: *Communication et langages* 53: 11–32 10.3406/colan.1982.1484.

Schöch, Christof (2016): „Does Short Sell Better? Belgian Author George Simenon’s use of sentence length“, in: *The Dragonfly’s Gaze* https://dragonfly.hypotheses.org/922 / http://dragonfly.hypotheses.org/1005 .

Spitzer, Leo ([1928]): „Die klassische Dämpfung in Racines Stil“, in: *Romanische Stil- und Literaturstudien* I. Marburg: Elwert (1931) 135–268.

Zur polykubistischen Informationsvisualisierung von Biographiedaten

Windhager, Florian
florian.windhager@donau-uni.ac.at
Donau-Universität Krems, Österreich

Mayr, Eva
eva.mayr@donau-uni.ac.at
Donau-Universität Krems, Österreich

Schreder, Günther
guenther.schreder@donau-uni.ac.at
Donau-Universität Krems, Österreich

Wandl-Vogt, Eveline
eveline.wandl-vogt@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Wien

Gruber, Christine
christine.gruber@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Wien

Methoden der Informationsvisualisierung dienen der Unterstützung menschlicher Kognition im Umgang mit abstrakten Daten und Themen (Scaife & Rogers, 1996). Dank der erfolgreichen Entwicklung entsprechender Verfahren helfen interaktive visuelle Repräsentationen seit geraumer Zeit bei der Analyse von multidimensionalen Daten in verschiedensten Disziplinen, inklusive zahlreicher geistes- und kulturwissenschaftlicher Forschungsfelder (cf. Sula, 2013; Jänicke, Franzini, Cheema & Scheuermann, 2015). Als Resultat ist mittlerweile ein ganzes Spektrum von bildgebenden Methoden für die Exploration und Analyse der Daten von geisteswissenschaftlichen ForscherInnen verfügbar. Dies gilt auch für HistorikerInnen, die biographische Datensätze von historischen Individuen exemplarisch mit Hilfe von geographischen Karten, chronologischen Timelines, genealogischen Bäumen, oder in relationalen Topologien und Netzwerken von Akteuren und Artefakten veranschaulichen und visuell analysieren können.

Der Vortrag baut auf dieser Vielheit von etablierten Methoden für die visuelle Analyse biographischer Daten auf – und präsentiert ein neue Methode der visuellen Synthese und Integration in einem konsistenten Rahmenwerk. Damit wird ein Vorschlag für die Gestaltung eines visuellen Interface unterbreitet, das die bessere kognitive Integration von mehreren möglichen Perspektiven auf komplexe historische Datensätze ermöglicht.

Das Bezugsproblem stellt dabei die kognitive Herausforderung dar, die auftritt, wenn multiple Visualisierungen (z.B. Karten, Treemaps oder Netzwerke – mit ihrer jeweiligen zeitlich-dynamischen Dimension) als Teilperspektiven auf denselben Datensatz zusammenkommen. Indem die resultierenden Bilder üblicherweise nur zeitlich gestaffelt (sequentiell) oder in räumlichem Nebeneinander (parallel, gelegentlich auch als „coordinated multiple views“, Roberts, 2007) präsentiert werden, stellt die makrokognitive Synthese (Klein & Hofmann, 2008) dieser lokalen Teilperspektiven zu einem globalen *bigger picture* eine besondere Herausforderung dar, das die kognitiven Systeme von ForscherInnen nicht selten überlastet. Kognitionswissenschaftlichen Reflexionen zum Gebrauch visueller Interfaces (Hegarty, 2011; Liu, Nersessian, & Stasko, 2008; Patterson et al., 2014) gehen davon aus, dass solche Synthesen qualitativ sehr unterschiedliche Ergebnisse zeitigen können – und dass ohne besonderen makrokognitiven Aufwand nur das Zustandekommen von unvollständigen und oftmals inkonsistenten „kognitiven Collagen“ (Tversky, 1993) zu erwarten ist.

Im Kontrast dazu präsentiert der Vortrag ein polykubistisches Rahmenwerk (Windhager, 2013; Windhager et al., 2016), das eine Synthese von unterschiedlichen Visualisierungen schon auf der Ebene der externen Repräsentation (i.e. des Displays) vornimmt, und somit die Konstruktion eines konsistenten mentalen Modells als interne Repräsentation erleichtert. Als grundlegende Methode der Visualisierung dienen hierbei sogenannte Raum-Zeit-Kuben (Space-Time Cubes), die zweidimensionale Visualisierungen (z.B. Karten) mit einer Zeitachse in der dritten Dimension zusammenführen. Die geographische Bewegung von Individuen oder Objekten wird in solchen Kuben als Raum-Zeit-Spur mit jeweils spezifischer und charakteristischer Gestalt sichtbar: Während ruhende Objekte vertikale Trajektorien in die Raumzeit zeichnen, werden Wanderungen und Ortsveränderungen als horizontale Abweichungen sichtbar, die in der Folge visuell analysiert werden können. Durch die freie Skalierbarkeit solcher Kuben können räumliche Bewegungen (von lokalen bis zu globalen Mustern) in allen zeitlichen Maßstäben (von Stunden bis zu Epochen) abgedeckt werden.

Dieses Verfahren, dass die beiden Visualisierungsmethoden von geographischen Karten und chronographischen Timelines zur Synthese bringt, kann in der Folge auf andere Methoden wie Treemaps oder Netzwerkvisualisierungen übertragen werden (Federico, Aigner, Miksch, Windhager, & Zenk, 2011; Windhager, 2013). Damit werden komplementäre Perspektiven auf die Lebenswege von Individuen durch die dreidimensionalen Topologien von geografischer, sozialer oder kulturell-kategorialer Raumzeit zusammengeführt (Abb. 1). Dieses Rahmenwerk von „coordinated multiple cubes“ dient durch seine generalisierte Projektionsmethode für zeit-orientierte Daten insofern zugleich der visuellen Analyse, sowie der visuellen Synthese von üblicherweise getrennten Einzelperspektiven. Die Kuben können in der Folge mit verschiedenen Methoden der dynamischen Visualisierung im Detail exploriert werden (Bach et al., 2014), sowie durch die Nutzung weiterer visueller Kohärenztechniken (z.B. narrative Methoden, cf. Windhager, Schreder, Smuc, & Mayr 2015) verwoben werden. Darüber hinaus wird durch die skizzierte Architektur die Trennung von Methoden der “Scientific Visualization” und der “Information Visualization” (Rhyne, 2003) überbrückt, wodurch Vorteile und Synergien für beide Seiten zum Tragen kommen (Sedlmair et al., 2009).

Um die praktische Relevanz dieses Rahmenwerks für die Exploration historischer Daten zu demonstrieren, präsentieren wir erste Ergebnisse der geo-temporalen Visualisierung von biographischen Datensätzen aus dem APIS-Projekt (http://www.oeaw.ac.at/acdh/en/apis) mithilfe der Software GeoTime (Kapler & Wright, 2005). In der Gegenüberstellung der Lebenswege von Individuen verschiedener Berufsgruppen (z.B. von Abenteurern und Kunstschaffenden) kommen strukturelle Merkmale zum Vorschein, die der visuellen Analyse komplexer historischer Datensätze neue Möglichkeiten eröffnen.

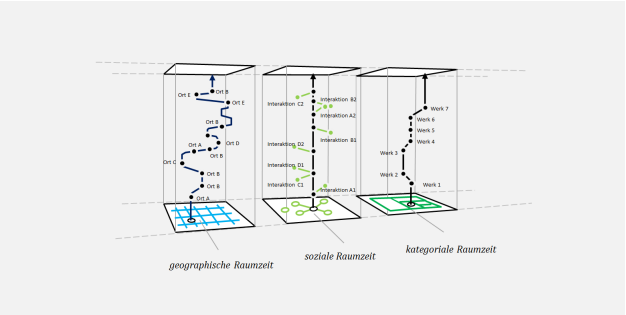


Abbildung 1: Rahmenwerk zur Visualisierung von Biographiedaten mit paralleler Perspektive auf geographische, soziale und kategoriale Raumzeit.

Bibliographie

Bach, Benjamin / Dragicevic, Pierre / Archambault, Daniel / Hurter, Christophe / Carpendale, Sheelagh (2014): „A Review of Temporal Data Visualizations Based on Space-Time Cube Operations“, in: *EuroVis-STARs*. The Eurographics Association 23–41.

Engelhardt, Yuri (2006): „Objects and spaces: The visual language of graphics“, in: *International Conference on Theory and Application of Diagrams*. Berlin / Heidelberg: Springer 104–108.

Federico, Paolo / Aigner, Wolfgang / Miksch, Silvia / Windhager, Florian / Zenk, Lukas (2011): „A Visual Analytics Approach to Dynamic Social Networks“, in: *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW), Special Track on Theory and Applications of Visual Analytics (TAVA)*. Graz: ACM 47:1–47:8.

Hegarty, Mary (2011): „The cognitive science of visual-spatial displays: Implications for design“ , in: *Topics in Cognitive Science* 3: 446–474.

Jänicke, Stefan / Franzini, Greta / Cheema, Muhammad Faisal / Scheuermann, Gerik (2015): „On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges“, in: *EuroVis-STARs*. The Eurographics Association.

Kapler, Thomas / Wright, William (2005): „GeoTime information visualization“, in: *Information Visualization* 4 (2), 136–146.

Klein, Gary / Hoffman, Robert R. (2008): „Macrocognition, mental models, and cognitive task analysis methodology“, in: *Naturalistic Decision Making and Macrocognition* 57–80.

Liu, Zhicheng / Nersessian, Nancy J. / Stasko, John T. (2008): „Distributed cognition as a theoretical framework for information visualization“, in: *IEEE Transactions on Visualization and Computer Graphics* 14 (6) 1173–1180.

Patterson, Robert E. / Blaha, Leslie M. / Grinstein, Georges G. / Liggett, Kristen K. / Kaveney, David E. / Sheldon, Kathleen C. / Moore, Jason A. (2014): „A human cognition framework for information visualization“, in: *Computers & Graphics* 42: 42–58.

Roberts, Jonathan C. (2007): „State of the art: Coordinated & multiple views in exploratory visualization“, in: *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV'07)* 61–71. IEEE.

Rhyne, Theresa-Marie (2003): „Does the difference between information and scientific visualization really matter?“, in: *IEEE Computer Graphics and Applications* 23 (3): 6–8.

Scaife, Mike / Rogers, Yvonne (1996): „External cognition: how do graphical representations work?“, in: *International Journal of Human-Computer Studies* 45 (2): 185–213.

Sedlmair, Michael / Ruhland, Kerstin / Hennecke, Fabian / Butz, Andreas / Bioletti, Susan / O'Sullivan, Carol (2009): „Towards the big picture: Enriching 3d models with information visualisation and vice versa“, in: *Smart Graphics*. Springer 27–39.

Sula, Chris Alen (2013): „Quantifying Culture: Four Types of Value in Visualisation“, in: Bowen, Jonathan P. / Keene, Suzanne / Ng, Kia (eds.): *Electronic Visualisation in Arts and Culture*. Springer 25–37.

Swaab, Roderick I. / Postmes, Tom / Neijens, Peter / Kiers, Marius H. / Dumay, Adrie C. M. (2002): „Multiparty negotiation support: The role of visualization's influence on the development of shared mental models“, in: **Journal of Management Information Systems** 19 (1): 129–150.

Tversky, Barbara (1993): „Cognitive maps, cognitive collages, and spatial mental models“, in: *Spatial Information Theory: A Theoretical Basis for GIS*. Berlin: Springer 14–24.

Windhager, Florian (2013): „On Polycubism. Outlining a Dynamic Information Visualization Framework for the Humanities and Social Sciences“, in: Fuellsack, Manfred (ed.): *Networking Networks: Origins, Applications, Experiments*. Wien: Turia + Kant 26–63.

Windhager, Florian / Mayr, Eva / Schreder, Günther / Smuc, Michael / Federico, Paolo / Miksch, Silvia (2016): „Reframing Cultural Heritage Collections in a Visualization Framework of Space-Time Cubes“, in: *Proceedings of the 3rd International Workshop on Computational History (HistoInformatics 2016)*. CEUR 20–24.

Windhager, Florian / Schreder, Günther / Smuc, Michael / Mayr, Eva (2015): „Drawing Things Together: Supporting Information Visualizations' Coherence across Multiple Views“, in: *Proceedings of the IEEE Information Visualization Conference 2016 (Posters Compendium)*. IEEE Computer Society Press.

Poster

AGATE – European Academies Internet Gateway: Konzept für eine digitale Infrastruktur für die geistes- und sozialwissenschaftlichen Forschungsvorhaben der europäischen Wissenschaftsakademien

Wuttke, Ulrike
wuttke@akademienunion-berlin.de
Union der deutschen Akademien der Wissenschaften, Deutschland

Adrian, Dominik
adrian@akademienunion-berlin.de
Union der deutschen Akademien der Wissenschaften, Deutschland

Ott, Carolin
ott@akademienunion-berlin.de
Union der deutschen Akademien der Wissenschaften, Deutschland

AGATE ist ein von der Union der deutschen Akademien der Wissenschaften (Akademienunion) koordiniertes Forschungsprojekt, das in enger Zusammenarbeit mit ALLEA, dem Zusammenschluss von mehr als 50 europäischen Akademien der Wissenschaften, durchgeführt wird. Die Union der deutschen Akademien der Wissenschaften ist die Dachorganisation von acht deutschen Wissenschaftsakademien. Ihre Hauptaufgabe ist die Koordination des Akademienprogramms, dem derzeit größten geisteswissenschaftlichen Forschungsprogramm in Deutschland. Bei der Mehrheit der geförderten Projekte handelt es sich um Langzeitvorhaben im Bereich der geisteswissenschaftlichen, aber auch der sozialwissenschaftlichen Grundlagenforschung.

Gefördert vom Bundesministerium für Bildung und Forschung ist das Projektziel das Ausloten des inhaltlichen, organisatorischen und technischen Rahmens für ein europäisches Akademienportal für die Geistes- und Sozialwissenschaften (European Academies Internet Gateway, kurz AGATE). Im Rahmen von AGATE sollen zum einen Informationen zu den umfangreichen geistes- und sozialwissenschaftlichen Forschungsaktivitäten an den europäischen Wissenschaftsakademien gebündelt zur Verfügung gestellt und die digitalen Forschungsergebnisse und -daten besser auffindbar und zugänglich gemacht werden. Zum anderen sollen Informationen zu nachhaltigen digitalen Forschungsmethoden und Publikationspraktiken bereitgestellt bzw. auf bereits bestehende Informations- und Serviceangebote verwiesen werden. Um diese Ziele zu erreichen, sieht der momentane Stand der Planung für die Plattform zwei grundlegende Komponenten mit verschiedenen Ausbaustufen vor: eine Projektedatenbank und ein so genannter *Service and Information Hub*.

Hintergrund

Die Grundidee für AGATE beruht auf den Erkenntnissen der SASSH-Umfrage (*Survey and Analysis of Basic Social Science and Humanities Research at the Science Academies and Related Research Organisations of Europe*, 2013-2015), in der erstmals über 600 Forschungsvorhaben an europäischen Wissenschaftsakademien und ähnlichen Forschungsinstitutionen systematisch zu verschiedenen Themengebieten befragt wurden (Leathem & Adrian 2015). Viele der europäischen Wissenschaftsakademien sind wichtige nationale Forschungszentren im Bereich der Geistes- und Sozialwissenschaften (SSH). Die Umfrage zeigte, dass die geistes- und sozialwissenschaftliche Forschung an den europäischen Wissenschaftsakademien angesichts der zunehmenden Digitalisierung mit großen Herausforderungen konfrontiert ist. Es zeigten sich bislang ungenutzte Potentiale in den Bereichen Kooperationen und Erfahrungsaustausch, digitale Infrastrukturen sowie digitale Forschungsmethoden und Publikationspraktiken.

Aus der Studie ging zum einen hervor, dass Kooperationen bzw. der Erfahrungsaustausch mit Forschungsvorhaben an anderen Akademien oftmals an mangelnden Informationen über potentielle Partnervorhaben scheitern. Zum anderen zeigte sich ein Nachholbedarf bezüglich des Wissensstandes über den Auftrag und die Angebote bzw. Kooperationsmöglichkeiten mit den europäischen SSH-Infrastrukturen (wie zum Beispiel CLARIN, DARIAH und Europeana). Des Weiteren wurde festgestellt, dass die geistes- und sozialwissenschaftliche Forschung der

Akademien im Internet und für die breitere Öffentlichkeit kaum sichtbar ist, wobei neben der stärkeren Nutzung des Internets als Kommunikationsweg über die Vorhaben und Projekte besonders eine stärkere Umsetzung von Prinzipien wie Open Access und Open Data die Verbreitung, Sichtbarkeit und Nachnutzung der digitalen Forschungsergebnisse erhöhen würde.

Während des AGATE Kick-Off-Workshops am 13. Juni 2016, bei dem unter anderem Vertreter verschiedener europäischer Wissenschaftsakademien Einblicke in die Herausforderungen, verfügbaren Lösungen und Desiderata im Bereich der geistes- und sozialwissenschaftlichen Akademienvorhaben durch den Digital Turn gaben, wurde wiederholt die Sicherung der Nachhaltigkeit der digitalen Forschungsmethoden und Publikationsmethoden als große Herausforderung betont.Mehr Informationen zum Programm des Workshops, einschließlich eines ausführlichen Berichts (Wuttke, Ott & Adrian, 2016), finden sich auf der AGATE-Projektseite. Durch die lange Dauer der von Akademien durchgeführten Forschungsvorhaben und die entsprechende langfristige Relevanz der Forschungsergebnisse spielen gerade in diesem Bereich ein intensiver, möglichst interdisziplinärer Wissensaustausch, die verstärkte Abstimmung und Bündelung der Aktivitäten und Ressourcen der Akademien untereinander und die Zusammenarbeit mit starken Infrastrukturpartnern eine wichtige Rolle. Generell würde hier eine bessere Zusammenarbeit mit europäischen Infrastrukturanbietern und -initiativen wie CLARIN, DARIAH, Europeana und OpenAIRE, beziehungsweise die verstärkte Nutzung ihrer Angebote und das Aufzeigen von Bedarfen zu einer Situation mit Gewinn für alle Beteiligten führen.

Entwicklung eines konzeptionellen Exposés für AGATE

Aufbauend auf den Erkenntnissen aus der SASSH-Umfrage und dem ersten Workshop sowie aus Expertengesprächen und Nutzerinterviews zeichnen sich momentan drei Schwerpunktbereiche heraus, die durch die Hauptkomponenten der im Rahmen des Posters vorgestellten paneuropäischen digitalen Infrastruktur für die geistes- und sozialwissenschaftliche Forschung (AGATE) adressiert werden sollten:

1) Sichtbarkeit und Konnektivität,

2) Wiederverwendung digitaler Projektergebnisse,

3) Nachhaltige digitale Forschungs- und Publikationspraktiken.

Die Ausarbeitung des konzeptionellen Exposés für AGATE ist von dem Grundgedanken getragen, wo immer möglich, auf bestehenden Angeboten aufzubauen, diese breiter bekannt zu machen und den Bedürfnissen der Wissenschaftlerinnen und Wissenschaftler der Akademien anzupassen, um somit die Zusammenarbeit zwischen den Akademien und den relevanten Infrastrukturen zu stärken.

1) Sichtbarkeit und Konnektivität

Trotz ihrer großen Bedeutung für die jeweilige nationale Wissenschaftslandschaft und ihrer langen Tradition, die sich insbesondere in der Langfristigkeit ihrer Forschungsvorhaben niederschlägt, sind Informationen über die an den europäischen Akademien durchgeführten geistes- und sozialwissenschaftlichen Forschungsprojekte in vielen Fällen schwer online auffindbar. Um die Sichtbarkeit und Konnektivität der geistes- und sozialwissenschaftlichen Akademienforschung zu verbessern, soll eine Projektedatenbank aufgebaut werden. Diese Datenbank würde so entwickelt und eingerichtet werden, dass sie detaillierte Informationen über die Forschungsaktivitäten der an Akademien angesiedelten Projekte und Vorhaben aufnehmen kann, wobei nicht nur klassische fachwissenschaftliche Kategorien (wie Forschungsgegenstand, Epoche, etc.), sondern auch digitale Methoden und Formate berücksichtigt würden, um den Wissensaustausch in diesen Bereichen zu befördern.

Die Datenbank würde sowohl Fachleuten als auch der interessierten Öffentlichkeit als zentrale Informationsquelle auf europäischer Ebene dienen. Gleichzeitig wäre sie für die Akademien ein einfaches und verhältnismäßig niedrigschwelliges Angebot, um die grundlegenden Informationen über ein Forschungsprojekt zu präsentieren, ohne eine eigene Projektwebseite aufbauen zu müssen.

Aus konzeptioneller und technischer Sicht stellt sich die Herausforderung, einen Katalog zu entwickeln, der es ermöglicht, Projekte nach einer Reihe relevanter Bereiche und Informationen zu erfassen, durchsuchen und zu clustern, und gleichzeitig möglichst intuitiv bedienbar ist. Zusätzlich soll ein Maximum an Konnektivität und Nachnutzung der Daten gewährleistet werden. Aus organisatorischer Sicht stellt sich die Frage, wie möglichst viele Projekte dazu bewegt werden können, sich in der Datenbank zu registrieren, bzw. die notwendigen Informationen zur Verfügung zu stellen.

2) Wiederverwendung digitaler Projektergebnisse

Um die Wiederverwendung digitaler Projektergebnisse durch bessere Auffindbarkeit zu steigern, soll die Datenbank von Beginn an so angelegt werden, dass in einem weiteren Schritt die verfügbaren digitalen Ressourcen der Akademien und Projekte aufgezeigt werden können. Unter den Begriff ‚digitale Ressource' (siehe u.a. Sahle 2015: 44) werden im Projektkontext sowohl digitale Publikationen in ‚klassischen' Formaten wie Artikel oder Monografien gefasst, als auch die in der Akademienforschung verbreiteten *enhanced publications* (wie Datenbanken, digitale Editionen und Wörterbücher), insbesondere *Work in Progress*, ebenso digitale ‚Quellen' wie Digitalisate oder Transkriptionen. Auch andere Formen wie Software-Code für DH-Tools sind denkbar. Der Anspruch an die Tiefe der Verknüpfung und Erschließung der digitalen Ressourcen beschränkt sich zunächst auf eine möglichst automatisierte Suche über Schnittstellen auf Metadatenebene und die weitergehende Betrachtung bzw. Forschung mit den ermittelten Ressourcen in ihrer originären Umgebung. Auch hier soll die Entwicklung in enger Abstimmung mit bestehenden infrastrukturellen Lösungen im europäischen Rahmen, wie etwa OpenAIRE, geschehen. AGATE würde somit den Weg ebnen, um einen zentralen Sucheinstieg für die heterogenen und verteilten digitalen geistes- und sozialwissenschaftlichen Ressourcen der europäischen Akademien zu entwickeln. Die Verknüpfung der heterogenen Ressourcen und digitalen 'Silos' würde die AGATE-Datenbank zu einem wertvollen Rechercheinstrument machen und einem breiten Publikum einen einfachen Zugang zu den digitalen Forschungsergebnisse der Akademien ermöglichen. Publikationen in Formaten, die für die geistes- und sozialwissenschaftliche Forschung an den Akademien typisch sind, wie Editionen, Wörterbücher und Korpora, könnten besonders hervorgehoben werden und würden dadurch erstmalig eine Plattform erhalten.

3) Nachhaltige digitale Forschungsmethoden und Publikationspraktiken

Um die generelle Stärkung der Nachhaltigkeit der digitalen Forschung durch wissenschaftlichen Erfahrungsaustausch und Kooperationen zwischen den Einzelakademien und darüber hinaus zu erreichen, insbesondere mit relevanten Infrastrukturpartnern und -initiativen auf nationaler, disziplinspezifischer und internationaler Ebene, ist ein so genannter Service and Information Hub als weitere Komponente von AGATE angedacht.

AGATE würde eine enge transnationale Zusammenarbeit und Kooperation unterstützen, indem Informationen über relevante Infrastrukturpartner, andere Organisationen und Initiativen, ihre Angebote und Kooperationsmöglichkeiten entweder durch aktive Mitwirkung oder als Datenlieferant bereitgestellt werden. AGATE würde auch ein umfangreiches Angebot an Informationen zu Schulungen sowie Materialien bereitstellen, die sich auf digitale Forschungs- und Publikationspraktiken beziehen (z. B. Werkzeuge, Standards, Richtlinien und Best Practices). Um Dopplungen zu vermeiden, würden die konkreten Informationsmodule, die auf dem Portal angeboten werden, in enger Zusammenarbeit mit den relevanten europäischen Infrastrukturen im Bereich der Geistes- und Sozialwissenschaften abgestimmt. Der Fokus läge vor allem darauf, auf einschlägige Ressourcen und Aktivitäten von Dritten zu verweisen (z.B. DiRT Directory, DHd-Blog), nicht eigene Materialien zu entwickeln.

In einem weiteren Schritt würden redaktionell betreute Informationen zu Spezialthemen wie Open Access oder Forschungsdatenmanagement im *Service and Information Hub* einen Platz bekommen, wobei der Schwerpunkt auf der Bewusstseinsförderung und praktischen Handreichungen läge. Des Weiteren könnten in diesem Rahmen digitale Forschungswerkzeuge sowie weitere im Kontext der Akademien entwickelte digitale Lösungen präsentiert werden (z.B. als „Tool des Monats"). Der *Service and Information Hub* würde ferner den Aufbau der Projektedatenbank flankieren, insbesondere wenn in diesem Rahmen konkrete Unterstützung für die Integration von Projektdaten und ggf. digitale Ressourcen in die Datenbank angeboten werden würde.

Der Service and Information Hub würde einerseits ein Forum für Erfahrungsaustausch und Kooperationen unter den IT- und Digital Humanities-Experten der europäischen Wissenschaftsakademien über digitale Tools und Methoden schaffen. Er würde aber auch andererseits eine Brücke zwischen dieser Community und den Fachwissenschaftlern schlagen und letztere stärker für Themen wie Open Access und Forschungsdatenmanagement sensibilisieren und aktiv befähigen.

Auf dem Poster werden Details der geplanten digitalen Infrastruktur vorgestellt. Dabei zeichnet sich beim bisherigen Stand der Arbeiten ab, dass bei der Konzeptionierung nicht nur innovative technische Lösungen

ausschlaggebend sind. Ebenso wichtig sind die Sicherstellung der organisatorischen Nachhaltigkeit der geplanten Infrastruktur, die größtmögliche Einbindung der wichtigsten Nutzergruppen schon in der Aufbauphase des Portals sowie rechtliche Fragen.

Das Projekt wird vom Bundesministerium für Bildung und Forschung (BMBF) unter dem Projekttitel „Aufbau eines europäischen Akademienportals“ (Laufzeit Oktober 2015-März 2017, Förderkennzeichen 01UG1503) gefördert.

Bibliographie

Akademienunion: http://www.akademienunion.de/ [letzter Zugriff 30. November 2016].

AGATE-Projektseite: http://www.akademienunion.de/agate/ [letzter Zugriff 30. November 2016].

ALLEA (ALL European Academies): www.allea.org [letzter Zugriff 30. November 2016].

Leathem, Camilla / Adrian, Dominik (2015): *Bestandsaufnahme und Analyse geistes- und sozialwissenschaftlicher Grundlagenforschung an den europäischen Wissenschaftsakademien und ähnlichen Forschungseinrichtungen*. Union der deutschen Akademien https://edoc.bbaw.de/files/1902/2015Projektpublikation_SASSH_deutsch_A1b.pdf [letzter Zugriff 30. November 2016].

Sahle, Patrick (2015): „Forschungsdaten in den Geisteswissenschaften“, in: *Bulletin SAGW* 4/2015: 43–45.

Wuttke, Ulrike / Ott, Carolin / Adrian, Dominik (2016): *AGATE: Chances and Challenges of a European Academies Internet Gateway: Kick-Off Workshop of the project “Elaboration of a Concept for a European Academies Internet Gateway (AGATE).* Workshop Report 1. Union of the German Academies of Sciences and Humanities http://www.akademienunion.de/fileadmin/redaktion/user_upload/Publikationen/BMBF-Projekt/AGATE_Erster_Workshop_Bericht_23.08.2016.pdf [letzter Zugriff 30. November 2016].

APIS – Eine Linked Open Data basierte Datamining-Webapplikation für das Auswerten biographischer Daten

Schlögl, Matthias
Matthias.Schloegl@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Lejtovicz, Katalin
katalin.lejtovicz@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Einführung

Das ÖBL (Österreichisches Biographisches Lexikon 1815-1950) ist ein umfassendes Werk, das derzeit rund 18,000 Biographien von wichtigen historischen Persönlichkeiten aus der österreichisch-ungarischen Monarchie und der Ersten und Zweiten Republik Österreichs enthält. Während an dem Lexikon noch gearbeitet wird, erscheint es in gedruckter Form, und seit 2009 ist es auch online verfügbar.
APIS - Mapping historical networks: Building the new Austrian Prosopographical | Biographical Information System - ist ein interdisziplinäres Digital Humanities Projekt, das WissenschaftlerInnen aus unterschiedlichen Themenbereichen (Biografien, Geschichte, Geographie, Sozialwissenschaften, Informationstechnologie) verbesserten Zugriff (Suchabfragen, API etc.) auf die ÖBL-Daten erlauben wird. Dadurch wird es möglich sein innovative, interdisziplinäre Forschung auf der Grundlage dieser einzigartigen Ressource durchzuführen. Als erstes Beispiel für eine solche angewandte wissenschaftliche Forschung und als wichtiger Test der Brauchbarkeit und Eignung der entwickelten Lösung, wird bereits im APIS Projekt eine soziodemografische Analyse, die die Formen und Muster der Migration von gesellschaftlichen Eliten untersucht, umgesetzt.

In unserer Präsentation konzentrieren wir uns auf die zugrunde liegende technische Lösung, vor allem auf die dynamischen Aspekte - Workflow – und die Ergebnisse der verschiedenen angewandten Verfahren, um den aktuellen Stand der Umsetzung zu beschreiben.

Ansatz

ÖBL Daten stehen momentan in einem Ad-hoc-XML-Format zur Verfügung. Diese XMLs enthalten einige Fakten (Geburts- und Todesdaten, Orte, Berufsangaben usw.) in strukturierter Form, der Großteil der Information versteckt sich jedoch in dem unstrukturierten Haupttext der Biographie. Das Hauptziel des Projektes ist Informationen automatisch aus dem freien Text zu extrahieren, und sie in strukturierter Form zur Verfügung zu stellen. Um dieses Ziel zu erreichen, wird ein zweifacher Hybrid-Ansatz verfolgt, der einerseits automatische und manuelle Textverarbeitung kombiniert und andererseits erlaubt die erhobenen Daten in verschiedenen Formaten zu serialisieren. Letzteres beinhaltet nicht nur die Bereitstellung in verschiedenen Formaten (z.B. RDF/JSON), sondern auch die Verwendung verschiedener Ontologien (z.B. CIDOC-CRM (Doerr 2003: 75-92), NDB (Historische Kommission bei der Bayerischen Akademie der Wissenschaften 1953)). Die extrahierten Entitäten sind mit mehreren semantischen Referenz Ressourcen wie zum Beispiel GND (Pfeifer 2012: 80-91), GeoNames [1] oder DBpedia (Bizer 2009: 154-165) abgeglichen und mit URIs aus diesen versehen (Entity Linking). Dieser kombinierte Ansatz wurde gewählt, um die höchstmögliche Genauigkeit der Annotationen zu gewährleisten, und den manuellen Aufwand so gering wie möglich zu halten. Obwohl es bewährte Techniken und Methoden für die Verarbeitung natürlicher Sprache gibt, wird manuelle Arbeit (Korrektur) der Forscher, die mit den jeweiligen Wissenschaftsgebieten vertraut sind, nach wie vor erforderlich sein.

Datenmodell

Das Datenmodell besteht aus fünf Entitäten (Personen, Institutionen, Orte, Werke und Ereignisse) und einer Meta-Entität (Verweis auf den ursprünglichen Artikel). Es gibt Beziehungen zwischen allen Entitäten (z.B. Person - Institution, Person - Ereignis) und Beziehungen sind auch zwischen den gleichen Objekttypen möglich (z.B. Person -> Vater_von -> Person). Die Beziehungen können auch temporalisiert (Start- und Enddatum) und typisiert werden (Typen können je nach Bedarf angegeben werden). Das erlaubt uns praktisch alle möglichen Szenarien zu modellieren.
Der ursprüngliche Plan war, die Daten nach bestehenden, gut definierten Ontologien zu modellieren. In der Evaluierungsphase wurde uns aber klar, dass sehr viele verschiedene Ontologien existieren. Einige sind wie CIDOC-CRM Event basiert, andere verbinden Entitäten direkt. Wir haben uns deshalb entschlossen ein eigenes (internes) Datenmodell zu erstellen und so den technischen Aufwand für die Verarbeitung, Darstellung und Speicherung der Daten möglichst gering zu halten. Gleichzeitig werden wir aber dieses interne Datenmodell mit Hilfe schon existierender Ontologien (NDB, CIDOC-CRM etc.) in verschiedenen Formen serialisieren und der Öffentlichkeit zur Verfügung stellen. Das stellt die möglichst einfache, nachhaltige Nutzung unserer Daten sicher.

Extraktion

Um strukturierte semantische Informationen aus den Biographien zu extrahieren, und die dadurch identifizierten Objekte zu Ressourcen wie GND, GeoNames zu verknüpfen verwenden wir automatische Tools. Die Ergebnisse werden von Experten verifiziert und ausgebessert um die Qualität der Daten zu gewährleisten, und um unser System durch manuelle Korrektur zu verbessern. Während die NLP-Tools eine schnelle Verarbeitung ermöglichen sind die Ergebnisse nicht zu 100% korrekt. Um die Genauigkeit zu verbessern, setzen wir mehrere Systeme, Quellen und Analysen ein.
Für die automatische Extraktion haben wir mehrere Tools getestet und bewertet, wie z.B. Stanford NER (Finkel 2005: 363-370), GATE (Cunningham 2011), OpenNLP [2], Stanbol (Bachmann-Gmur 2013), basierend auf folgende Kriterien: 1) welche Sprachen unterstützt das System 2) Möglichkeit der Anpassung, 3) Entity Linking Fähigkeiten, 4) Output Format und 5) die Verfügbarkeit und Qualität der API. Apache Stanbol hat sich als das am besten geeignete Werkzeug für unsere Zwecke gezeigt. Stanbol ermöglicht die Verknüpfung von Entitäten wie Personen, Institutionen zu Referenzressourcen (Normdateien, Ontologien). Wir haben die Biographien mit GND und

GeoNames abgeglichen, und planen weitere LOD [3] Ressourcen hinzuzufügen. Durch die Verknüpfung von oben benannten Entitäten zu den semantischen Ressourcen können wir viele zusätzliche Informationen (z.B. Alternative Namen, Titel von Werken usw.) zu unseren Daten hinzufügen, und so Inhalte mit fehlenden Informationen bereichern.

Anwendung

Um den manuellen Arbeitsaufwand (Korrektur der Daten etc.) zu minimieren haben wir eine effiziente und einfache Weboberfläche geschaffen, die es den ForscherInnen erlaubt mit den Daten zu interagieren. Im Sinne einer nachhaltigen Nutzung und einfacher weiteren Betreuung des so entstandenen Tools haben wir uns entschlossen auf erprobte Web-Technologien zu setzen (Django [4] /MySQL). Die Web-Anwendung ist in Django, einem Python-basierten Web-Entwicklungs-Framework, implementiert. Django ist nicht nur ein ausgereiftes und verbreitetes Tool (Websites wie Disqus, Pinterest und die Washington Times nutzen es), sondern bietet auch die Möglichkeit die volle Bandbreite der verschiedenen Python Bibliotheken nativ im Code zu verwenden (NLTK [5] , scikit-learn [6] , NumPy [7] etc.) .
Die Web-Anwendung stellt die Daten der einzelnen Biographien strukturiert in drei Teilen dar: primäre minimale Informationen, Haupttext mit markierten Anmerkungen und die Listen von Orten, Institutionen und Personen, die mit dem Biographierten in Zusammenhang stehen. Die Anwendung bietet auch Funktionen für die Navigation: dropdown Listen sowie einfache Volltextsuche.
Eine weitere wichtige Funktion der Anwendung ist die Möglichkeit, den Text manuell mit Annotationen zu versehen. Dieses Feature erlaubt sowohl die Korrektur von automatischen Annotationen, als auch das Hinzufügen von neuen Annotationen. Die Kuratoren können die Entitäten mit der Maus auswählen oder im Kontextmenü identifizieren.

Derzeit liegt der Schwerpunkt auf der Darstellung von Orten. Dementsprechend wird die Anwendung mit eingebetteten Karten ausgestattet, an denen identifizierte geographische Orte visualisiert werden können. In der nächsten Phase des Projekts wird eine interaktive Visualisierung entwickelt, um das Verständnis der Daten und die Navigation im Datenbestand zu erleichtern.

Arbeitsablauf

Das System unterstützt zwei Workflows: im ersten Schritt schickt die Anwendung (das Extrakt-Modul) die Biographien im Batch-Modus zu einem Extraktionsservice (lokale Stanbol Instanz), welches die Abfragen an externe Services und/oder lokale Indizes weiterleitet und die gematchten Entitäten in einer Liste in JSON-LD Format zurückgibt. Diese Entitäten werden von dem Extrakt-Modul analysiert und in der Datenbank abgespeichert. Danach werden sie in der Web-Anwendung dargestellt und können von den ForscherInnen überprüft und korrigiert werden.
Im zweiten Schritt wird der Workflow vom Benutzer gestartet: Der menschliche Annotator markiert einen String und identifiziert ihn als Ort, die Anwendung schickt den ausgewählten String zur Stanbol Instanz, die die verfügbaren Ressourcen abfragt und mögliche Kandidaten zurückgibt. Diese Treffer werden dem/der ForscherIn in Form eines Autocomplete Feldes angezeigt.

Schlussfolgerung

Während wir uns in unserem Abstrakt auf die technische Umsetzung konzentriert haben, ist es wichtig im Auge zu behalten, dass das System nur eine Voraussetzung ist die eigentliche Forschungsfragen beantworten zu können. Alle im Projekt generierten Daten sowie die entwickelte Forschungsumgebung wird der Öffentlichkeit zugänglich gemacht (eine erste Version der Forschungsumgebung wird Ende September in unserem Github Account zugänglich gemacht). Wie schon weiter oben angesprochen versuchen wir die Nachhaltigkeit unserer Lösung auf mehrfache Weise zu erreichen. Zum einen verwenden wir gut etablierte Web-Technologien und ermöglichen somit vielen Entwicklern weltweit unseren Code zu warten und/oder weiter zu entwickeln. Zum anderen verbinden wir unsere Daten mit der LOD-Cloud und serialisieren sie mit Hilfe verschiedener weit verbreiteter Ontologien in den gängigsten Formaten und stellen so sicher, dass andere Projekte unsere Daten mit äußerst kleinem Aufwand direkt in ihre Projekte einbetten können.

Fußnoten

1. http://www.geonames.org/

2. https://opennlp.apache.org/
3. http://linkeddata.org/
4. https://www.djangoproject.com/
5. http://www.nltk.org/
6. http://scikit-learn.org/stable/
7. http://www.numpy.org/

Bibliographie

APIS: Mapping historical networks: Building the new Austrian Prosopographical | Biographical Information System (APIS) http://www.oeaw.ac.at/acdh/en/apis

Bachmann-Gmur, Reto (2013): *Instant Apache Stanbol* (1st ed.). Packt Publishing. ISBN 1783281235.

Bizer, Christian / Lehmann, Jens / Kobilarov, Georgi / Auer, Soren / Becker, Christian / Cyganiak, Richard / Hellmann, Sebastian (2009): „DBpedia - A crystallization point for the Web of Data“, in: *Journal of Web Semantics* 7 (3): 154–165.

Cunningham, Hamish / Maynard, Diana / Bontcheva, Kalina (2011): *Text Processing with GATE* (Version 6). University of Sheffield Department of Computer Science. ISBN 0956599311.

Doerr, Martin (2003): „The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata“, in: *AI Magazine* 24 (3): 75–92.

Finkel, Jenny Rose / Grenager, Trond / Manning, Christopher (2005): „Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling“, in: *Proceedings of ACL-2005* 363–370.

Historische Kommission bei der Bayerischen Akademie der Wissenschaften (seit 1953): *Neue deutsche Biographie*, Berlin: Duncker & Humblot. ISBN 3-428-00181-8

ÖBL - Österreichisches Biographisches Lexikon/Austrian Biographical Lexicon (1815-1950) Online-Edition und Österreischiches Biographisches Lexikon ab 1815 (2. Überarbeitete Auflage - online). Verlag der Österreichischen Akademie der Wissenschaften. Wien. http://www.biographien.ac.at/oebl [letzter Zugriff 26. August 2016]

Pfeifer, Barbara (2012): „Vom Projekt zum Einsatz. Die gemeinsame Normdatei (GND)“, in: Brintzinger, Klaus-Rainer (ed.): *Bibliotheken: Tore zur Welt des Wissens*. 101. Deutscher Bibliothekartag in Hamburg 2012, Olms, Hildesheim u.a. 2013: 80–91.

Comparison of Methods for Automatic Relation Extraction in German Novels

Krug, Markus
markus.krug@uni-wuerzburg.de
Universität Würzburg, Deutschland

Wick, Christoph
christoph.wick@uni-wuerzburg.de
Universität Würzburg, Deutschland

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

Reger, Isabella
isabella.reger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Weimer, Lukas
lukas.weimer@uni-wuerzburg.de
Universität Würzburg, Deutschland

Madarasz, Nathalie
nathalie.madarasz@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

Puppe, Frank
frank.puppe@uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Die automatische Erkennung von spezifischen Relationen ermöglicht Einsichten über die Beziehungen zwischen Entitäten. Solche Informationen können nicht nur als Kantenbezeichner in sozialen Netzwerken fungieren, sondern auch als globale Constraints für das schwierige Problem der Coreference Resolution eingesetzt werden. Darüber hinaus kann eine Relationserkennung zur Beantwortung diverser literarischer Fragestellungen eingesetzt werden, z.B. ob eine Romangattung sich mit bestimmten Relationstypen befasst, oder ob die Arten

der Relationen sich über die Jahrhunderte verändern. In dieser Arbeit stellen wir ein Label-Set für die Extraktion von binären Relationen zwischen Personen-Entitäten vor und vergleichen Feature-basierte Ansätze des maschinellen Lernens mit regelbasierten Ansätzen zur automatischen Erkennung dieser Relationen. Da Trainingsmaterial zur Verfügung steht, liegt der Fokus in dieser Arbeit auf dem Einsatz überwachter Methoden, d.h. unsere regelbasierten Verfahren sind ebenfalls auf einer zuvor abgetrennten Menge entwickelt worden. Wir verwenden ein neues Korpus, das manuell mit mehr als 50 verschiedenen, hierarchisch gegliederten Relationstypen annotiert wurde.

Related Work

Eine Übersicht über Arbeiten zur Relationserkennung findet sich in [Jung et al. 2012] sowie [Bach und Badaskar 2007]. Sowohl für den überwachten, als auch den halb-überwachten Fall wurden erfolgreiche Methoden entwickelt. Da dieses Paper sich hauptsächlich auf überwachte Algorithmen bezieht, geben wir nur einen knappen Überblick über halb-überwachte Verfahren.

Algorithmen zur Relationsextraktion erhalten typischerweise zwei (oder mehr) Referenzen zu Entitäten (sogenannte Instanzen) als Input und sollen die Klasse, und das dazugehörige Label, vorhersagen, welche die Relation zwischen den Entitäten beschreibt. Die meisten Experimente wurden anhand englischer Texte und den Datensätzen der Automatic Content Extraction (ACE) Workshops 2004 und 2006 durchgeführt. Auf dem Datensatz von 2004 wurden Experimente zur Unterscheidung von 5 und 27 verschiedenen Klassen wie Arbeitsplatz-, körperliche, soziale, Mitgliedschafts- und Diskursrelationen (wobei manche Unterklassen von anderen sein können) betrachtet. Hierfür gibt es zahlreiche Ansätze, die jedoch alle versuchen, eine diskriminative Beschreibung der Instanzen zu erhalten und diese davon ausgehend zu klassifizieren:

- In der Feature-basierten Klassifikation wird eine Instanz (normalerweise zwei Referenzen zu Entitäten) durch einen Feature-Vektor mit manchmal mehr als einer Million Dimensionen repräsentiert und mit Methoden wie Maximum Entropy Modellen [Kambhatla 2004] oder Support Vector Machines [Jiang und Zhai 2007] klassifiziert. Der letztere Ansatz konnte auf den ACE2004-Daten einen F1-Score von 72,9% für die Erkennung von 7 verschiedenen Relationen erzielen. In unseren Experimenten verwenden wir für die Feature-basierten Methoden ähnliche Features wie Kambhatla [Kambhatla 2004].
- Kernel-basierte Klassifikation wurde häufig zur Relationsextraktion genutzt und liefert konkurrenzfähige Ergebnisse [Zhou et al. 2007, Zhang et al. 2006, Zhao und Grishman 2005]. Während Feature-basierte Verfahren die Instanz direkt repräsentieren, funktionieren Kernel-basierte Methoden etwas anders. Aus einer technischen Perspektive kann ein Kernel als eine Funktion betrachtet werden, die zwei Instanzen als Input erhält (also ein Paar von Referenzen) und direkt einen Wert berechnet, der auf der "Ähnlichkeit" dieser Instanzen basiert, wobei einer höherer Wert eine größere Ähnlichkeit anzeigt. Es wurden zahlreiche Kernel für die Relationsextraktion vorgeschlagen; eine tiefgehende Analyse und Erklärung findet sich in Jung et al. [Jung et al. 2012].
- Die regelbasierte Klassifikation verwendet eine für den Menschen lesbare Repräsentation durch Regeln, die entweder manuell erstellt oder gelernt wurden. Als Vorteile können die inhärente Erklärungsfähigkeit und die einfache Integration in Feature-basierte Machine Learning-Verfahren gesehen werden.

Im Folgenden vergleichen wir die genannten Methoden anhand eines Label-Sets zur Erkennung binärer Relationen zwischen Figuren in manuell annotierten Abschnitten von deutschsprachigen Romanen.

Annotation, Datensatz und Vorverarbeitung

Da Textstellen, an denen Relationen zwischen Entitäten explizit benannt werden, in Romanen typischerweise rar sind, ist es nicht sinnvoll, komplette Romane zu annotieren, da der Ertrag an Daten zu gering wäre. Aus diesem Grund wurde zunächst eine kleine Teilmenge per Hand annotiert und dann genutzt, um mit einem MaxEnt Classifier in einer Active Learning-Umgebung neue Sätze zum Labeln vorschlagen zu können. (Ein Überblick hierzu findet sich in Finn und Kushmerick [Finn und Kushmerick 2003]). Diese Umgebung erhielt Sätze aus 312 verschiedenen Romanen von Projekt Gutenberg

und 215 Zusammenfassungen aus dem Kindler Literatur Lexikon Online. Daraus entstand ein Korpus mit 2412 Sätzen, die insgesamt 1265 Relationen enthalten (was wiederum die Knappheit an Daten illustriert). 33 Texte wurden zufällig für die Testmenge ausgewählt, sodass es feste Test- und Trainingsdaten gibt (1988 respektive 424 Sätze mit 1070 respektive 195 Relationen). Die verwendeten Label sind ähnlich zu Massey et al. [Massey et al. 2015]. Die Relationen werden durch eine Ontologie mit momentan 57 verschiedenen Relationstypen repräsentiert, die hierarchisch geordnet sind (beispielsweise ist die Relation "Tochter" der Relation "Familie" untergeordnet). Abbildung 1 zeigt die oberste Ebene des Label-Sets, mit den gleichen Kategorien wie in Massey et al. [Massey et al. 2015] und einer zusätzlichen Relation "Liebe".

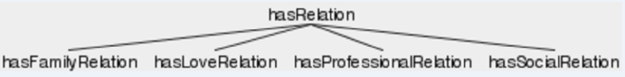


Abbildung 1: Die ersten beiden Ebenen unseres verwendeten Label-Sets mit den vier Haupttypen, die sich weiter in insgesamt 57 Relationstypen untergliedern lassen.

Eine Relation wurde von einem Annotator als ein benannter, gerichteter Bogen zwischen zwei Entitäten in einem Satz gelabelt, sofern sie explizit im Text beschrieben ist. Es wurde immer das spezifischste Label verwendet, da die übergeordneten Relationstypen (vgl. Abbildung 1) daraus abgeleitet werden können. Abbildung 2 zeigt ein Beispiel einer Relation, wie sie in unserem Korpus annotiert ist.

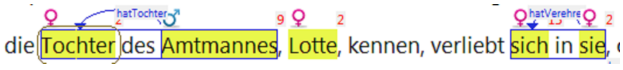


Abbildung 2: Zwei gelabelte Instanzen von Relationen in unserem Datensatz. Die erste zeigt die Relation “hatTochter” und die zweite die Relation “hatVerehrer”.

Um solche Relationen automatisch erkennen zu können, müssen die Texte eine große Zahl an Vorverarbeitungsschritten durchlaufen. Wir verwenden die Figurenerkennung von Jannidis et al. [Jannidis et al. 2015] und die gleiche Vorverarbeitung wie in [Krug et al. 2016].

Experimente

Wir verwenden einen regelbasierten Ansatz mit manuell erstellten Regeln und zwei Feature-basierte Lernverfahren (Maximum Entropy, MaxEnt und Support Vector Machines, SVM). Der regelbasierte Ansatz nutzt sowohl die textuelle Repräsentation, als auch den kürzesten Pfad im Dependency-Baum und formuliert die Regel auf Basis dieser Repräsentationen und der Repräsentationen aus dem reinen Text. Das folgende Beispiel zeigt Regeln, die zu den Relationen aus Abbildung 2 passen:

- Tochter des <Entität> => hatTochter(2,1)
- Pfad: <Entität>->verliebt<->in<-<Entität> => hatVerehrer(2,1)

Die erste Regel basiert auf der angepassten Text-Repräsentation, während die zweite Regel sich auf den kürzesten Dependency-Pfad zwischen "sich" und "sie" bezieht. Die Zahlen in runden Klammern geben die Richtung an (in beiden Fällen von Entität 2 auf Entität 1). Die Regeln wurden manuell auf den zuvor gewählten Trainingsdaten erzeugt. Insgesamt wurden fast 500 solcher Regeln ermittelt. Der Großteil der Relationen konnte jedoch mit 3 Regeln (ab hier sogenannte Core-Regeln) abgedeckt werden, die Possessiv- und Genitivkonstruktionen abbilden.

Die Feature-basierten Ansätze wurden in zwei Szenarien evaluiert: a) nur mit bereits bekannten Features aus Related Work und b) mit zusätzlichen Booleschen Features (eines pro Regel), falls eine der 500 Regeln passt.

Tabelle 1 zeigt die Evaluationsergebnisse der verschiedenen Methoden für drei hierarchische Ebenen (alle Relationen, Relationen der obersten Ebene, alle 57 Relationstypen) und Tabelle 2 die Ergebnisse für die vier Relationstypen der obersten Ebene. Während die Verwendung aller Regeln zu einem F1-Score von 71% für alle Relationen und 59% für die vier übergeordneten Relationstypen führt, erreicht der Feature-basierte Ansatz mit MaxEnt mit einem Booleschen Feature für jede Regel etwas bessere Ergebnisse (F1 von 73,6% und 61,2%). Ohne die Regel-Features liegt der Score der Lernverfahren deutlich niedriger. Die SVM erreicht teilweise eine höhere Precision als MaxEnt, aber im Allgemeinen einen signifikant geringeren F1-Wert.

Tabelle 1: Ergebnisse der verschiedenen Ansätze für drei verschiedene Evaluationsszenarien: binär (das reine Vorliegen einer Relation), für die 4 Haupttypen und für alle 57 Relationstypen insgesamt.

Tabelle 2: Ergebnisse für die verschiedenen Ansätze, aufgeschlüsselt nach den 4 Haupttypen. Familienrelationen erreichen sehr gute

Ergebnisse mit einem F1-Wert von fast 80% und einer Precision von bis zu 95%. Liebesrelationen sind schwerer zu erkennen, liegen aber dennoch bei 56,3% F1. Die anderen Relationstypen fallen in der Qualität ab, sind aber gleichzeitig weniger relevant.

Sehr auffällig ist das gute Ergebnis für die drei Core-Regeln und dabei besonders die hervorragende Precision von 96,2% für Familien-Relationen. Eine genauere Betrachtung der False Positives (FP) in Tabelle 3 zeigt, dass diese Relationen fast immer syntaktisch korrekt erkannt wurden, aber semantisch irrelevant und daher nicht im Goldstandard annotiert sind (z.B. "mein Gott"). Hier zeigt sich eine Schwachstelle dieser Arbeit: teilweise unpräzise Richtlinien für die Annotation von Relationen. Das ist jedoch ein sehr schwieriges Problem, das eventuell umgangen werden kann, wenn die Relationserkennung kein Ziel in sich, sondern eine untergeordnete Aufgabe im Zuge der Erkennung der Hauptfiguren und deren Beziehungen in Romanen ist.

Tabelle 3: Auswertung der drei Core-Regeln auf unserem Datensatz

Regel	Beispiel	TP	FP
<Possessive> ... <Entity>	seine liebe Mutter [his loved mother]	83	22
<Entity_Noun> <GENITIV_Noun>	Frau des Kanzlers [wife of the chancellor]	21	7
<GENITIV_Noun> <Entity_NN>	Peters Frau [Peter's wife]	19	5

Fazit und zukünftige Arbeiten

Dieses Paper hat gezeigt, dass automatische Relationserkennung eine Herausforderung darstellt. Einfache Regeln können jedoch bereits einen wesentlichen Teil der Relationen mit hoher Precision erkennen. Dennoch ist der Bedarf an weiteren Verbesserungen durch fortschrittliche Methoden hier deutlich. Zudem ist die Evaluation der Relationserkennung an sich schwierig und kann besser im Kontext eines übergeordneten Ziels wie der automatischen Erstellung eines Netzwerks der Hauptfiguren eines Romans [Krug 2016] oder der Gattungsklassifikation [Hettinger et al. 2015] eingebracht werden.

Bibliography

Bach, Nguyen / Badaskar, Sameer (2007): „A review of relation extraction“, in: *Literature review for Language and Statistics* II.

Finn, Aidan / Kushmerick, Nicolas (2003): „Active learning selection strategies for information extraction“, in: *Proceedings of the International Workshop on Adaptive Text Extraction and Mining (ATEM-03)*.

Hettinger, Lena / Becker, Martin / Reger, Isabella / Jannidis, Fotis / Hotho, Andreas (2015): „Genre classification on German novels“, in: *Proceedings of the 12th International Workshop on Text-based Information Retrieval*.

Jannidis, Fotis / Krug, Markus / Reger, Isabella / Toepfer, Martin / Weimer, Lukas / Puppe, Frank (2015): „Automatische Erkennung von Figuren in deutschsprachigen Romanen“, in: *DHd 2015: Von Daten zu Erkenntnissen*.

Jiang, Jing / Zhai, ChengXiang (2007): „A Systematic Exploration of the Feature Space for Relation Extraction“, in: *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*.

Jung, Hanmin / Choi, Sung-Pil / Lee, Seungwoo / Song, Sa-Kwang (2012): „Survey on Kernel-Based Relation Extraction“, in: Sakurai, Shigeaki (ed.): *Theory and Applications for Advanced Text Mining*. InTech Open Science 10.5772/51005.

Kambhatla, Nanda (2004): „Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations“, in: *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*.

Krug, Markus / Fotis, Jannidis / Reger, Isabella / Weimer, Lukas / Macharowsky, Luisa / Puppe, Frank (2016): „Attribuierung direkter Reden in deutschen Romanen des 18.-20. Jahrhunderts. Methoden zur Bestimmung des Sprechers und des Angesprochenen“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung*.

Krug, Markus / Fotis, Jannidis / Reger, Isabella / Weimer, Lukas / Macharowsky, Luisa / Puppe, Frank (2016): „Comparison of Methods for the Identification of Main Characters in German Novels“, in: *DH2016: Converence Abstracts*.

Massey, Philip / Xia, Patrick / Bamman, David / Smith, Noah A. (2015): „Annotating Character Relationships in Literary Texts“, in: *arXiv*, arXiv:1512.00728.

Zhao, Shubin / Grishman, Ralph (2005): „Extracting relations with integrated

information using kernel methods“, in: *Proceedings of ACL-2005*.

Die Odyssee zum richtigen Standard - Herausforderungen einer konsistenten Datenmigration von *Ulysses: A Critical and Synoptic Edition* (1984)

Schäuble, Joshua
joshua.schaeuble@uni-passau.de
Universität Passau, Deutschland;
Reichsuniversität Groningen, Niederlande

Crowley, Ronan
crowle01@gw.uni-passau.de
Universität Passau, Deutschland

Mit „Ulysses: A Critical and Synoptic Edition“ erschien 1984 eine der ersten Forschungseditionen, die auf Basis der systematischen Verwendung von Kollationierungssoftware digital erzeugt wurde. Das Münchner Team um Hans Walter Gabler verwendete hierzu TUSTEP sowohl zur Validierung der Transkripte einzelner Zeugen als auch zur Erschließung der zeugenübergreifenden Synopse. Für die gedruckte Edition wurden die halbautomatisch erzeugten Kollationsergebnisse mit einem eigens entwickelten System komplexer Diakritika ausgezeichnet, die es dem geübten Leser ermöglichen sollten, die Textentstehung über stellenweise mehr als zwanzig inter- und intradokumentarische Textstufen hinweg in einer synoptisch integrierten Textfassung nachzuvollziehen. Während die Konzeption und Umsetzung dieser Arbeit bis heute als bahnbrechend im Bereich der Computerphilologie zu bezeichnen ist, konnte das Potenzial der resultierenden Druckausgabe für die Joyce-Forschung nicht annähernd ausgeschöpft werden. Zu komplex war das Markup, dem es gelingen sollte, zu verknüpfen, was zuvor getrennt war und zu hoch war der Aufwand, sich in diese Systematik einzuarbeiten.

Im Digitalen hingegen führten die Daten jene Odyssee fort, die die Druckedition beenden sollte. Auf der Suche nach einem Markup-Standard, der es vermag, die Inhalte der Druckedition digital zu repräsentieren, wurden die TUSTEP Ergebnisse zunächst von Tobias Rischer im Rahmen seiner Diplomarbeit (1997) in SGML/TEI transformiert und anschließend in mehreren Überarbeitungen über TEI P4 bis hin zur aktuellen Version der TEI P5v3 (2016) migriert. Dieser Beitrag vollzieht die Evolution dieser “Legacy Data” nach, bis hin zu ihrer jüngsten Station - der noch andauernden Bemühung einer Migration nach TEI P5v3, welche im Rahmen des DFG- und NEH-geförderten Kooperationsprojektes “Diachronic Markup and Presentation Practices for Text Editions in Digital Research Environments” am Lehrstuhl für Digital Humanities der Universität Passau durchgeführt wird.

Erstmals seit der zweiten, überarbeiteten Ausgabe der synoptisch-kritischen Gabler Edition 1986 gelang es, aus den TEI-Daten die synoptische Visualisierung der Druckedition zu rekonstruieren und somit eine Konsistenzprüfung gegen die ursprünglichen Daten zu ermöglichen. Erst durch diese visuelle Rückführung offenbarten sich migrationsbedingte Fehler und Provisorien, welche zuvor, wenn überhaupt, nur in Fußnoten und privaten Aufzeichnungen vergangener Beteiligter dokumentiert wurden. Neben dem allgemeinen Versuch, die vollzogenen Änderungen aus den Aufzeichnungen und Migrationsergebnissen früherer Projekte zu rekonstruieren, hat es sich das Passauer Team zur Aufgabe gemacht, Strategien zur Entdeckung, Typisierung und Korrektur derartiger „Migrationsverluste“ zu entwickeln. Ein wesentlicher Bestandteil dieser Arbeit ist die Abschätzung der Leistungsfähigkeit und Wirtschaftlichkeit von automatisierten Batch-Konvertierungen mittels XSLT und Python im Vergleich zur manuellen Intervention und Korrektur der Kodierung.

Neben der Identifikation und Korrektur von „Migrationsfehlern“, steht die Rekonstruktion der textgenetischen Perspektive, durch welche sich die Druckedition auszeichnete, im Vordergrund. Während Gabler die textuelle Entwicklung, welche er mittels der Kollation chronologisch aufeinander folgender Textzeugen erschlossen hatte, im Druck synoptisch darstellen konnte, beinhalteten die TEI Guidelines bis zur Version P5v2 kein Modell zur Auszeichnung textgenetischer Prozesse. Es fehlte schlicht die Möglichkeit zur formalisierten Dokumentation einer stufenweisen,

zeugenübergreifenden Chronologie der Textentwicklung. In der Druckedition wurde jeder auktorialen Textänderung *genau eine* Textstufe aus der heuristisch erschlossenen Chronologie zugeordnet. Diese lineare Textentwicklung über intra- und interdokumentarische Textstufen, in Gablers Terminologie auch Overlay und Level genannt, musste im Digitalen in eine Auszeichnung überführt werden, welche die Genese in den Hintergrund rückt und zu jeder auktorialen Modifikation anstelle einer Textstufe eine Liste sämtlicher Zeugen verzeichnet, auf welcher die spezifische Änderung Bestand hat. Diese Art der dokumentenorientierten Kodierung von Textgenese entspricht zwar bis heute der gängigen Auszeichnungspraxis historisch-kritischer Editionsprojekte, repräsentierte aber zu keinem Zeitpunkt die textgenetische Intension der 84er *Ulysses* Edition. Erst mit der Integration eines textgenetischen Modells in die TEI Guidelines, kann die ursprüngliche Intension erstmals auch in TEI kodiert werden. Hierzu bedarf es einer weiteren Episode der Datenmigration auf der Odyssee zum richtigen Standard.

Bibliographie

Brüning, Gerrit / Henzel, Katrin / Pravida, Dietmar (2014): „Multiple Encoding in Genetic Editions: The Case of Faust“, in: *Journal of the Text Encoding Initiative* 4. Available from: jtei.revues.org.

Burnard, Lou / O'Brien O'Keeffe, Katherine / Unsworth, John (2006): *Electronic Textual Editing*. New York: Modern Language Association of America.

Burnard, Lou / Jannidis, Fotis / Pierazzo, Elena / Midell, Gregor / Rehbein, Malte (2010): „An Encoding Model for Genetic Editions“, in: *TEI: Text Encoding Initiative*. Retrieved from www.tei-c.org/ Activities/Council/Working/ tcw19.html/.

Joyce, James / Gabler, Hans Walter (eds.) (1984): *Ulysses: A Critical and Synoptic Edition*. New York: Garland.

Joyce, James (1922): *Ulysses*. Paris: Shakespeare and Company.

Fordham, Finn (2010): *I do, I undo, I redo: The Textual Genesis of Modernist Selves in Hopkins, Yeats, Conrad, Forster, Joyce, and Woolf*. Oxford / New York: Oxford University Press.

Rischer, Tobias (1997): *Eine TEI/SGML-Edition der textkritischen Ausgabe von James Joyces Ulysses*. Diplomarbeit, LMU München.

TEI Consortium (eds.) (2016): *TEI P5: Guidelines for Electronic Text Encoding and Interchange. P5v3*. Available from: http:// www.tei-c.org/Guidelines/P5/ .

Digitale Erschließung einer Sammlung von Volksliedern aus dem deutschsprachigen Raum

Burghardt, Manuel
manuel.burghardt@ur.de
Lehrstuhl Medieninformatik, Universität Regensburg, Deutschland

Spanner, Sebastian
sebastian.spanner@stud.uni-regensburg.de
Lehrstuhl Medieninformatik, Universität Regensburg, Deutschland

Schmidt, Thomas
thomas.schmidt@stud.uni-regensburg.de
Lehrstuhl Medieninformatik, Universität Regensburg, Deutschland

Fuchs, Florian
florian.fuchs@stud.uni-regensburg.de
Lehrstuhl Medieninformatik, Universität Regensburg, Deutschland

Buchhop, Katia
katia.buchhop@stud.uni-regensburg.de
Lehrstuhl Medieninformatik, Universität Regensburg, Deutschland

Nickl, Miriam
miriam.nickl@stud.uni-regensburg.de
Lehrstuhl Medieninformatik, Universität Regensburg, Deutschland

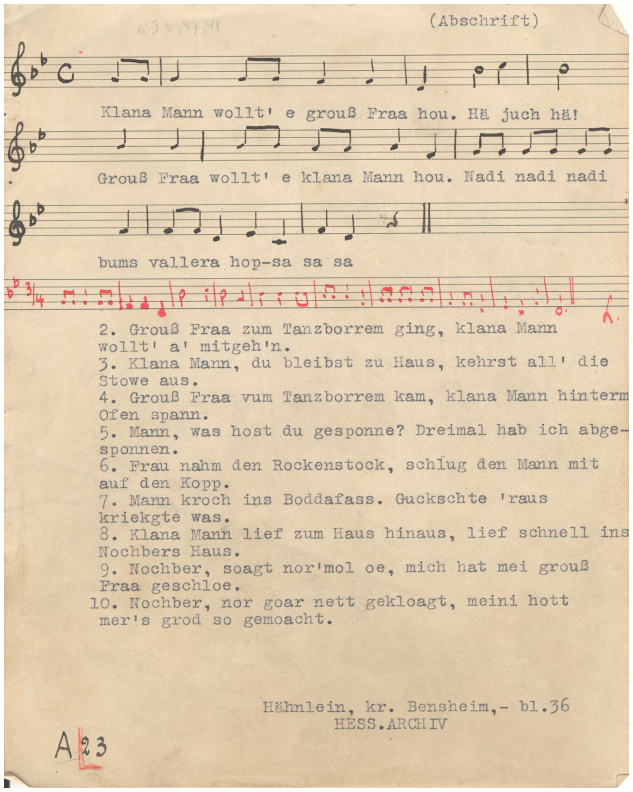
Wolff, Christian
christian.wolff@ur.de
Lehrstuhl Medieninformatik, Universität Regensburg, Deutschland

Projektkontext

Dieser Beitrag beschreibt ein laufendes Projekt [1] zur digitalen Erschließung einer großen Sammlung von Volksliedern aus dem deutschsprachigen Raum, mit dem Ziel diese später über ein öffentliches Informationssystem verfügbar zu machen. Mithilfe dieses Informationssystems soll neben der üblichen Exploration gescannter Faksimiles der Originalliedblätter zusätzlich ein quantitativer Zugang zu den Daten ermöglicht werden, der diese anhand unterschiedlicher Parameter durchsuchbar und analysierbar macht. Ziel des Projekts ist also nicht nur, einen in dieser Form einzigartigen Bestand an Liedblättern nachhaltig digital zu erschließen und zugänglich zu machen, sondern darüber hinaus computergestützt nach Auffälligkeiten in Form wiederkehrender Phrasen und Themen oder melodischen Universalien zu suchen, die für verschiedene Regionen oder Zeitabschnitte charakteristisch sind.

Datenbasis

Die Datengrundlage des Projekts stellen umfangreichen Quellen zur Volksmusikforschung dar, die seit einigen Jahren von der Universitätsbibliothek Regensburg verwaltet werden. Die Regensburger Liedblattsammlung umfasst etwa 140.000 Blätter mündlich oder handschriftlich tradierter Volkslieder aus dem gesamten deutschsprachigen Raum, und ist, was Abdeckung und Umfang angeht, in dieser Form einzigartig (Krüger, 2013). Die losen Einzelblätter enthalten einerseits handschriftliche, monophone Melodien und andererseits Liedtexte, welche zumeist mit Schreibmaschine verfasst wurden (vgl. Abb. 1).


(Abschrift)

Klana Mann wollt' e grouß Fraa hou. Hä juch hä!

Grouß Fraa wollt' e klana Mann hou. Nadi nadi nadi

bums vallera hop-sa sa sa

2. Grouß Fraa zum Tanzborrem ging, klana Mann wollt' a' mitgeh'n.
3. Klana Mann, du bleibst zu Haus, kehrst all' die Stowe aus.
4. Grouß Fraa vum Tanzborrem kam, klana Mann hinterm Ofen spann.
5. Mann, was host du gesponne? Dreimal hab ich abgesponnen.
6. Frau nahm den Rockenstock, schlug den Mann mit auf den Kopp.
7. Mann kroch ins Boddafass. Guckschte 'raus kriekgte was.
8. Klana Mann lief zum Haus hinaus, lief schnell ins Nochbers Haus.
9. Nochber, soagt nor'mol oe, mich hat mei grouß Fraa geschloe.
10. Nochber, nor goar nett gekloagt, meini hott mer's grod so gemoacht.

Hähnlein, kr. Bensheim,- bl.36
HESS.ARCHIV

A 23

Abbildung 1: Ausschnitt aus dem Liedblatt Nr. A23: „Klana Mann wollt' e grouß Fraa hou“.

Zu den Liedblättern existieren darüber hinaus Metadaten wie *Titel*, *Text-Incipit*, *Sangesort* und *Jahr*, die ursprünglich in einem umfangreichen Zettelkastensystem vorlagen, mittlerweile jedoch in eine Datenbank (*Augias*) übertragen wurden. In Zusammenarbeit mit der Universitätsbibliothek Regensburg werden zunächst Scans der Liedblätter erstellt und mit den bereits vorhandenen digitalen Metadaten verknüpft. Daraufhin werden die Scans inhaltlich erfasst und in ein maschinenlesbares Format gebracht, das erlaubt, die Daten computergestützt zu durchsuchen und zu analysieren. Dieser Beitrag beschreibt Herausforderungen und Lösungsansätze bei der digitalen Erschließung der Liedblätter hinsichtlich ihrer Texte und Melodien.

Digitale Erschließung der Liedblätter

Für die Transkription der Texte und Melodien wurden Tools für die automatische Erfassung evaluiert. Neben automatischer Texterkennung (OCR, *Optical Character Recognition*), wurde auch die automatische Notenerkennung (OMR, *Optical Music Recognition*) untersucht (vgl. Bainbridge &

Bell, 2001; 2006; Raphael & Wang, 2011; Rebelo, Capela, & Cardoso, 2010).

Erschließung der Liedtexte über OCR mit manueller Nachkorrektur

Die Evaluation der Eignung bestehender OCR-Tools für den Kontext der Regensburger Liedblattsammlung lehnt sich an Kanungo, Marton und Bulbul (1999) an. Das Testkorpus umfasst 102 Liedblätter, die möglichst viele unterschiedliche typographische und orthographische Phänomene abdecken, etwa Druckschrift (mit unterschiedlich starkem Kontrast), Frakturschrift, aufgeklebte Korrekturen, Sonderzeichen, etc. Für die Evaluation wurde die Textzone unterhalb der Notenzeilen ausgewählt, da die Noten als unbekannte Sonderzeichen das Texterkennungsergebnis negativ verfälschen würden. Für jene Textzonen wurde eine manuelle Transkription erstellt, die in der weiteren Evaluation als *ground truth* dient. Evaluiert wurden die folgenden drei OCR-Tools:

- *Abbyy Fine Reader* (http://www.abbyy.de/)
- *Omnipage Professional (*http://www.nuance.de/for-individuals/by-product/omnipage/index.htm)
- *Adobe Acrobat X Pro* (https://helpx.adobe.com/de/acrobat/kb/acrobat-downloads.html)

Mithilfe des OCR-Evaluationstools *ocrevalUAtion* (Carassco, 2014) wurde jeweils der Output der drei getesteten OCR-Tools mit den *ground truth*-Daten verglichen. Abb. 2 zeigt für jedes OCR-Tool die Anzahl korrekt erkannter Zeichen (*correct*), die Anzahl falsch erkannte Zeichen (*confused*), die Anzahl nicht erkannte Zeichen (*lost*) sowie die Anzahl überflüssiger Zeichen (*spurious*) als gestapeltes Balkendiagramm.

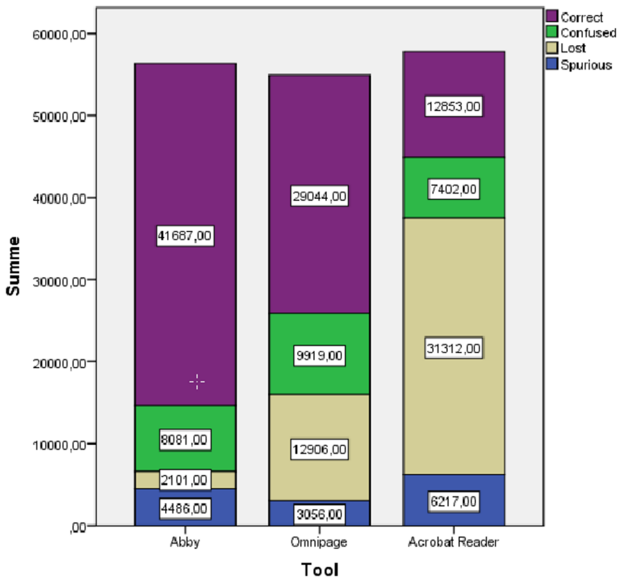


Abbildung 2: OCR-Evaluationsergebnisse für die getesteten Tools hinsichtlich der korrekt erkannten, der falsch erkannten, der gar nicht erkannten sowie der überflüssigerweise erkannten Zeichen.

Anhand dieser Parameter lassen sich Kennzahlen für die Tools berechnen, etwa die *precision* oder auch die global *error rate*. Bezüglich der korrekten Erkennung in Prozent wird deutlich, dass *Abbyy* mit einer Erkennungsrate von 80% (*Omnipage*: 56%, *Adobe*: 26%) und einer vergleichsweise geringen Streuung am besten in der Evaluation abschneidet (vgl. Abb. 3).

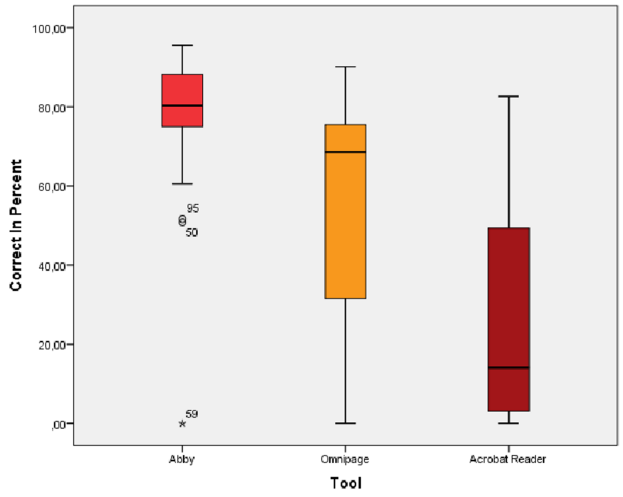


Abbildung 3: Boxplot zur Erkennungsgenauigkeit der einzelnen OCR-Tools.

Dass *Abbyy*-Tool liefert die besten Evaluationsergebnisse und wurde somit als OCR-Tool für die Liedblattsammlung ausgewählt. Die 80%-Erkennungsrate erlaubt erste explorative

Analysen der Liedblätter anhand bestimmter Schlüsselwörter. Für die sukzessive Korrektur der Texte wurde ein Tool entwickelt, das die manuelle Korrektur des OCR-Outputs für jedes Liedblatt erlaubt. Um die Texte der insgesamt 140.000 Liedblätter möglichst effizient zu transkribieren, sind zudem weitere Evaluationsexperimente mit anderen OCR-Tools geplant. Zudem soll versucht werden, das *Abbyy*-Tool anhand der Liedblätter zu trainieren, um so die Erkennungsrate weiter zu verbessern.

Erschließung der Melodien über ein Crowdsourcing-Webtool

In Anlehnung an eine OMR-Evaluationsstudie (Bellini, Bruno & Nesi, 2007) wurden drei der am weitesten verbreiteten OMR-Tools hinsichtlich ihrer Eignung für die Liedblattsammlung evaluiert:

- *Photoscore* (http://www.sibelius.com/products/photoscore/ultimate.html)
- *SharpEye* (https://www.columbussoft.de/SharpEye.php)
- *CapellaScan* (http://www.capella.de/de/index.cfm/produkte/capella-scan/info-capella-scan/)

Anders als bei der OCR-Evaluation ist die Erstellung eines automatisch abgleichbaren *ground truth*-Datensatzes nicht ohne weiteres möglich, da die Erfassung musikalischer Notation wesentlich komplexer ist als reine Textzeichenerkennung. Der Abgleich des jeweiligen OMR-Outputs mit dem entsprechenden Originalliedblatt erfolgte deshalb manuell. Insgesamt wurden auf diese Weise 20 Liedblätter ausgewählt, welche eine möglichst hohe Bandbreite unterschiedlicher Merkmalsausprägungen abdecken. Zu den Merkmalen zählen Zeichenabstand, Einheitlichkeit der Zeichen, allgemeiner Kontrast, Kontrast der Notenlinien, Größe der Notenköpfe, Länge der Notenhälse und das Vorkommen von Fremdzeichen.

Bei der Berechnung der Erkennungsgenauigkeit wurden dieselben Parameter verwendet wie schon bei der OCR-Evaluation (vgl. Abb. 2). Die Ergebnisse der OMR-Evaluation zeigen, dass hinsichtlich der durchschnittlichen Erkennungsgenauigkeit mit 36% bei *Photoscore*, 8% bei *CapellaScan* und 4% *SharpEye* keines der Tools auch nur ansatzweise für den produktiven Einsatz in Frage kommt (vgl. Abb. 4). Dabei ist selbst beim am besten evaluierten Tool *Photoscore* eine enorme Streuung zu beobachten, die bei 5 von 20 Blättern auf 0% kommt, und nur ein einziges Mal als beste Erkennungsrate 80% bei einem Liedblatt erreicht.

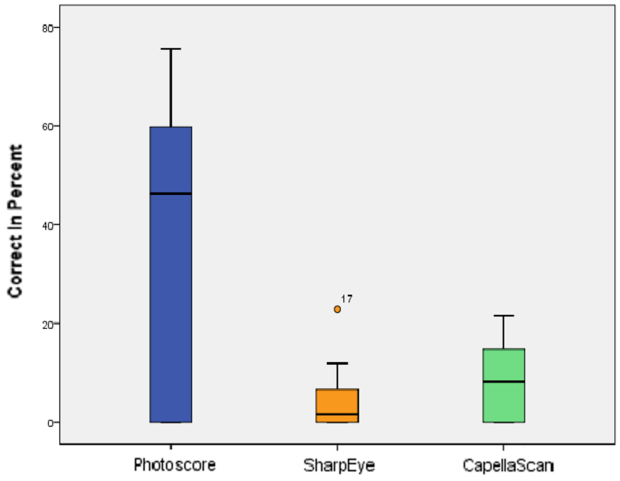


Abbildung 4: Boxplot zur Erkennungsgenauigkeit der einzelnen OMR-Tools.

Als alternative Erschließungsstrategie wurde ein Transkriptionstool namens *Allegro* entwickelt, welches aufgrund der erheblichen Datenmenge von mehreren tausend Liedblättern auf einen Crowdsourcing-Ansatz (Dunn & Hedges, 2013; Oomen & Aroyo, 2011) zurückgreifen soll. Erfolgreiche Beispiele für solche Ansätze im Bereich der Digital Humanities finden sich etwa beim Sammeln und Dokumentieren von urbaner Kunst (Burghardt, Schneider, Bogatzki, & Wolff, 2015), bei der Transkription von Manuskripten (Causer & Wallace, 2012), bei der Verschlagwortung von Kunstwerken (Commare, 2011) und auch im Bereich der Transkription von Musikstücken, wie beim Projekt „What's the Score?“ [2].

Bei der Umsetzung des Tools für die Transkription der Regensburger Liedblätter wurde besonderes Augenmerk auf die einfache Bedienbarkeit durch iteratives *usability testing* während des Entwicklungsprozesses gelegt (vgl. ISO 13407:1999). Die Benutzeroberfläche wurde dabei so konzipiert, dass auch Personen, die keine Noten lesen können, in der Lage sind, die Noten zu transkribieren, indem sie diese auf ein virtuelles Notenblatt übertragen und das Original im Wesentlichen nachbauen (vgl. Meier et al., 2015). Die zusätzliche Möglichkeit der Transkription über ein Midi-Instrument

soll später über einen speziell anzuwählenden Expertenmodus optional verfügbar gemacht werden.

Als erster Schritt wird in *Allegro* zunächst das Notenblatt manuell in einzelne Takte segmentiert (Abb. 5):

Abbildung 5: Taktweise Segmentierung der Liedblätter mit dem *Allegro*.

Nach Angabe der Liedblattnummer sowie der Auswahl von Taktart und Tonart gelangt man in den eigentlichen Transkriptionsmodus, bei dem Takt für Takt auf einer interaktiven Notenzeile mit Maus und Tastatur (Shortcuts) transkribiert wird (vgl. Abb. 6). Jeder einzelne Takt kann im Browser abgespielt werden, um so ggf. auf auditiver Ebene schnell Transkriptionsfehler zu erkennen.

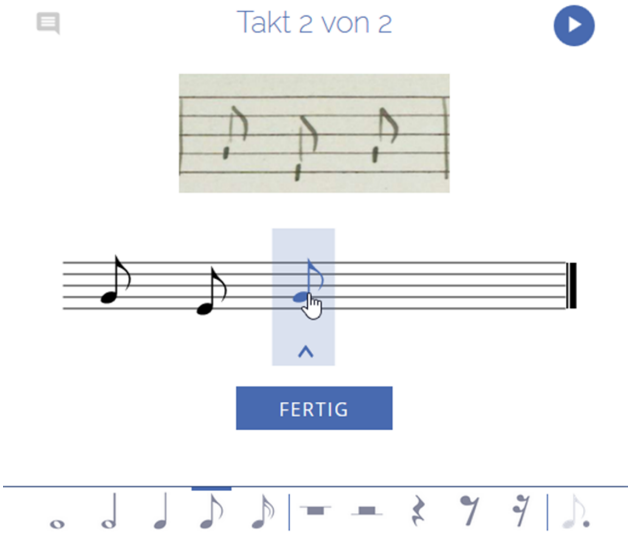


Abbildung 6: Taktweise Transkription der Liedblätter mit dem Allegro-Tool.

Im Hintergrund werden die Eingaben auf das virtuelle Notenblatt schließlich in ein maschinenlesbares Format (*JSON*) übersetzt, das mithilfe einer *Converter*-Toolbox in beliebige andere Formate wie etwa *MusicXML* transformiert werden kann. Da die Transkription durch Laien eine erhöhte Gefahr für Transkriptionsfehler mit sich bringt, wird jedes Liedblatt doppelt übersetzt (vgl. das *double keying*-Konzept bei Texttranskriptionen). Liedblätter, bei denen die Transkriptionen nicht übereinstimmen, werden auf redaktioneller Ebene final geprüft. Um den Anreiz zur Beteiligung an der Transkription zu erhöhen, ist es den Teilnehmern möglich die selbst transkribierten Texte und Melodien in einer privaten Sammlung zu speichern und bei Bedarf als PDF bzw. als MP3 herunterzuladen.

Das Transkriptionstool befindet sich aktuell in der offenen Beta-Testphase und findet guten Zuspruch bei den Anwendern:

- *Allegro*: http://allegro.sytes.net/

Zusammenfassung

Dieser Beitrag gibt einen Einblick in ein laufendes Projekt zur digitalen Erschließung einer großen Sammlung von Liedblättern. Während OCR-Tools für die automatische Erfassung der Liedtexte annehmbare Ergebnisse mit einer Erkennungsrate von bis zu 80% liefern, so liegt die Erkennungsgenauigkeit bestehender OMR-Tools für die handschriftlichen Notensätze bei lediglich maximal 36%. Im Falle der Notenerkennung wurde von Grund auf ein neues, intuitiv bedienbares Transkriptionstool entwickelt, welches über einen Crowdsourcing-Ansatz die sukzessive Erschließung der Notensätze sicherstellen soll.

Ausblick

Aktuell liegt der Projektfokus auf der Erschließung der Liedblätter. Parallel entstehen zudem erste Prototypen (vgl. Burghardt et al., 2016) für das angedachte Informationssystem, das die Analyse der Liedblätter anhand der verfügbaren Metadaten, der Liedtexte sowie anhand verschiedener melodischer Parameter (vgl. Mongeau & Sankoff, 1990; Orio & Rodá, 2009; Typke, 2007) erlaubt. Im Rahmen des weiteren Projektverlaufs sollen anhand der digital erschlossenen Liedblätter u.a. die folgenden Fragestellungen untersucht werden:

- Welche sind die häufigsten Wörter in den Texten deutscher Volkslieder, und welche Wörter treten besonders häufig zusammen auf (Kollokationen)? Lassen sich daraus Rückschlüsse auf wiederkehrende Themen ziehen, einerseits für das gesamte Liedblattkorpus, andererseits aus einer regionalen und diachronen Perspektive?
- Gibt es melodische Universalien, die typisch für deutsche Volkslieder sind, einerseits für das gesamte Liedblattkorpus, andererseits aus einer regionalen und diachronen Perspektive?
- Lassen sich musikalisch-linguistische Kollokationen identifizieren, kommen also bestimmte Melodien oder einzelne Rhythmen

oder Intervalle besonders häufig in Texten mit auffälligen Schlüsselwörtern vor?

Fußnoten

1. Anmerkung: Erste Vorarbeiten zu den hier beschriebenen Vorhaben erfolgten im Rahmen des DFG-Projekts „Erschließung von Quellen der Volksmusikforschung, Zugänglichmachung durch Digitalisierung sowie virtuelle Wiederherstellung zerstreuter Bestände“, vgl. http://rvp.ur.de.
2. Projekt „What’s the Score?“ online: https://www.bodleian.ox.ac.uk/weston/our-work/projects/whats-the-score

Bibliographie

Bainbridge, David / Bell, Tim (2001): „The challenge of optical music recognition“, in: *Computers and the Humanities* 35: 95–121.

Bellini, Pierfranceso / Bruno, Ivan / Nesi, Paolo (2007): „Assessing Optical Music Recognition Tools“, in: *Computer Music Journal* 31 (1), 68–93.

Burghardt, Manuel / Lamm, Lukas / Lechler, David / Schneider, Matthias / Semmelmann, Tobias (2016): „Tool-based Identification of Melodic Patterns in MusicXML Documents“, in: *Digital Humanities 2016: Conference Abstracts* 440–442.

Burghardt, Manuel / Schneider, Patrick / Bogatzki, Christopher / Wolff, Christian (2015): „StreetartFinder – Eine Datenbank zur Dokumentation von Kunst im urbanen Raum“, in: *DHd 2015: Von Daten zu Erkenntnissen*.

Carrasco, Rafael C. (2014): „An open-source OCR evaluation tool“, in: *DATeCH 2014*. New York: ACM Press.

Causer, Tim / Wallace, Valerie (2012): „Building A Volunteer Community: Results and Findings from Transcribe Bentham“, in: *DHQ: Digital Humanities Quarterly* 6 (2).

Commare, Laura (2011): „Social Tagging als Methode zur Optimierung Kunsthistorischer Bilddatenbanken – Eine empirische Analyse des Artigo-Projekts“, in: *Kunstgeschichte. Open Peer Reviewed Journal* urn:nbn:de:bvb:355-kuge-160-9.

Dunn, Stuart / Hedges, Mark (2013): „Crowd-sourcing as a Component of Humanities Research Infrastructures“, in: *International Journal of Humanities and Arts Computing* 7 (1-2): 147–169.

Kanungo, Tapas / Marton, Gregory A. / Bulbul, Osama (1999): „Performance evaluation of two Arabic OCR products“, in: *The 27th AIPR workshop: Advances in computer-assisted recognition* 76–83.

Krüger, Gerd (2013): „Das ‚Regensburger Volksmusik-Portal‘ der Universitätsbibliothek Regensburg: Bestände – Problematiken – Perspektiven: Zwischenbericht aus einem Erschließungsprojekt“, in: Mohrmann, Ruth-E. (ed.), *Audioarchive – Tondokumente digitalisieren, erschließen und auswerten*. Münster et al.: Waxmann Verlag 119–131.

Meier, Florian / Bazo, Alexander / Burghardt, Manuel / Wolff, Christian (2015): „A Crowdsourced Encoding Approach for Handwritten Sheet Music“, in: *Music Encoding Conference Proceedings 2013 and 2014* 127–130.

Mongeau, Marcel / Sankoff, David (1990): „Comparison of Musical Sequences“, in: *Computers and the Humanities* 24: 161–175.

Oomen, Johan / Aroyo, Lora (2011): „Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges“, in: *C&T ’11 Proceedings of the 5th International Conference on Communities and Technologies* 138–149.

Orio, Nicola / Rodà, Antonio (2009): „A Measure of Melodic Similarity Based on a Graph Representation of the Music Structure“, in: *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)* 543–548.

Raphael, Christopher / Wang, Jingya (2011): „New Approaches to Optical Music Recognition“, in: *12th International Society for Music Information Retrieval Conference (ISMIR)* 305–310.

Rebelo, Ana / Capela, G. / Cardoso, Jaime S. (2010): „Optical recognition of music symbols“, *International Journal on Document Analysis and Recognition* 13: 19–31.

Typke, Rainer (2007): *Music Retrieval based on Melodic Similarity*. Ph.D Thesis, Utrecht University.

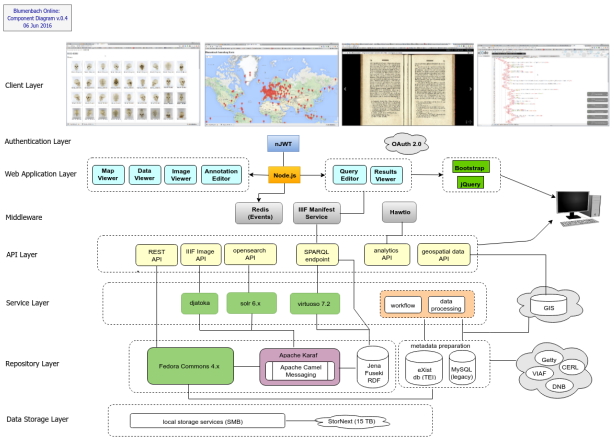
Digitale Nachhaltigkeit bei Grundlagenforschung in Akademieprogramm: Das Beispiel „Johann Friedrich Blumenbach-online“

Wettlaufer, Jörg
jwettla@gwdg.de
Akademie der Wissenschaften zu Göttingen, Deutschland

Johnson, Christopher
christopher.johnson@uni-goettingen.de
Akademie der Wissenschaften zu Göttingen, Deutschland

Das Projekt „Johann-Friedrich Blumenbach online“ ist ein Forschungsvorhaben der Akademie der Wissenschaften zu Göttingen mit einer Laufzeit von 15 Jahren bis 2024. Es hat sich zum Ziel gesetzt, die gedruckten Veröffentlichungen sowie die Sammlungstätigkeit dieses Göttinger Gelehrten und Begründers der physischen Anthropologie, die sich über einen längeren Zeitraum im letzten Viertel des 18. und im ersten Drittel des 19. Jahrhunderts erstreckt, durch eine digitale Edition für die wissenschaftshistorische Forschung zu erschließen. Eine besondere Herausforderung für dieses ganz digital konzipierte Projekt ist dabei der Aufbau einer nachhaltigen digitalen Infrastruktur über die gesamte Laufzeit des Projekts und darüber hinaus.

Nach einer Phase der Zusammenarbeit mit Textgrid, einer virtuellen Forschungsumgebung im Rahmen von dariah.de, wurde seit 2015 ein neuer Ansatz für eine nachhaltige Präsentation der Ergebnisse präferiert, der auf Linked Open Data (LOD) basiert und sich neben den transkribierten XML-Texten und den Metadaten der erfassten Objekte auf die digitalisierten Abbildungen von Texten und Objekten selber konzentriert. Diese sollen in einer PANDORA [Presentation (of) ANnotations (in a) Digital Object Repository Architecture] genannten Annotationsumgebung für die Forschung erschlossen werden. Technologisch und erkenntnistheoretisch schließt sich das Vorhaben den Standards an, die seit einigen Jahren vom International Image Interoperability Framework (IIIF) [1] vorangetrieben werden. Aufbauend auf die Schnittstellen dieses Frameworks bietet PANDORA über ein sog. „Manifest“ die Organisation der Präsentation von Bilddaten, die z.B. in einem Repository gespeichert werden. Dieses „Manifest“ besteht aus einem JSON-LD [2] Dokument und wird dynamisch aus einem digitalen Objektrepositorium mit Hilfe von SPARQL-Abfragen [3] erzeugt. Es orientiert sich dabei an der Semantik und dem Konzept der „IIIF Presentation API“ [4] . Die Architektur des Systems ist in mehreren Schichten organisiert, die jeweils unterschiedliche Funktionen abdecken. Die unterste Ebene bildet der Repository-Layer, in dem mit Fedora Commons eine mächtige Speicherlösung zur Verfügung steht, mit der die Bild- und Objektdaten sowie die zugehörigen Annotationen verwaltet werden können. In den darauf aufbauenden Schichten (Service, API und Web Application Layer) werden die Daten für die Annotationen und Präsentation aufbereitet und schließlich im Client Layer in einem Viewer visualisiert.



Mit der Verwendung aktueller standardisierter APIs, die von namhaften Einrichtungen der Kulturgutbewahrung eingesetzt und unterstützt werden, erhofft sich das Projekt eine besondere Nachhaltigkeit der Investitionen, die in den Aufbau des Portals und der Forschungsumgebung fließen. Durch eine Entkoppelung der Open-Source Komponenten in PANDORA, die bei Bedarf einfach ausgetauscht werden können, ohne die Grundfunktionalität zu gefährden, soll eine langfristige und ressourcenschonende Verwendung der Forschungsumgebung

gewährleistet werden. Für die Präsentation der Bilddaten können verschiedene Viewer wie z.B. mirador [5] eingesetzt werden, ohne dass eine spezielle Anpassung notwendig ist. Aufgrund der auf Semantic Web Technologien aufbauenden Architektur sowie der Möglichkeit der Bereitstellung der Daten als Linked Open Data über einen Triplestore (Jena Fuseki) ist eine Nachnutzungsmöglichkeit der Daten durch andere Projekte gegeben, die ebenfalls zur Nachhaltigkeit der PANDORA-Lösung beiträgt.

Langfristige Forschungsvorhaben im Akademienprogramm stehen vor besonderen Herausforderungen, da noch während der Projektlaufzeit technologische (Weiter-) Entwicklungen zu erwarten sind, die bei Antragstellung weder vorhergesehen noch vollständig antizipiert werden können. Dazu kommt die Herausforderung, digitale Systeme nicht nur für die Präsentation der Ergebnisse sondern auch für den Arbeitsprozess bei deren Erstellung vorzuhalten und kontinuierlich weiter zu entwickeln. Das Poster möchte, am Beispiel des Projekts „Johann-Friedrich Blumenbach online“, eine aktuelle und nachhaltige Lösung für diese Aufgabenstellung vorstellen und in der deutschen Digital Humanities Community diskutieren.

Fußnoten

1. http://iiif.io/
2. https://www.w3.org/TR/json-ld/
3. https://www.w3.org/TR/sparql11-query/
4. http://iiif.io/api/presentation/2.1/
5. http://github.com/IIIF/mirador

Bibliographie

Kerzel, Martina / Reich, Mike / Weber, Heiko (2013): „Die Edition ‚Johann Friedrich Blumenbach – online‘ der Akademie der Wissenschaften zu Göttingen“, in: Neuroth, Heike / Lossau, Norbert / Rapp, Andrea (eds.): *Evolution der Informationsinfrastruktur: Kooperation zwischen Bibliothek und Wissenschaft*. Glückstadt: Verlag Werner Hülsbusch 107–136.

Lauer, Gerhard (2014): „Johann Friedrich Blumenbach – online (Projektbericht für das Jahr 2013)“, in: *Jahrbuch der Akademie der Wissenschaften zu Göttingen: 2013*. Boston / Berlin: De Gruyter 235–237.

Wettlaufer, Jörg / Johnson, Christopher / Scholz, Martin / Fichtner, Mark / Thotempudi, Sree Ganesh (2015): „Semantic Blumenbach: Exploration of Text-Object Relationships with Semantic Web Technology in the History of Science“, in: Terras, Melissa / Clivaz, Clare / Verhoeven, Deb / Kaplan, Frederic (eds.): *Digital Scholarship in the Humanities* (DSH), Special Issue „Digital Humanities 2014“ 30, Supplement 1: i187–i198.

Digitale Nachhaltigkeit in den Geisteswissenschaften durch TOSCA: Nutzung eines standardbasierten Open-Source Ökosystems

Breitenbücher, Uwe
uwe.breitenbuecher@iaas.uni-stuttgart.de
Institut für Architektur von Anwendungssystemen, Universität Stuttgart, Deutschland

Barzen, Johanna
johanna.barzen@iaas.uni-stuttgart.de
Institut für Architektur von Anwendungssystemen, Universität Stuttgart, Deutschland

Falkenthal, Michael
michael.falkenthal@iaas.uni-stuttgart.de
Institut für Architektur von Anwendungssystemen, Universität Stuttgart, Deutschland

Leymann, Frank
frank.leymann@iaas.uni-stuttgart.de
Institut für Architektur von Anwendungssystemen, Universität Stuttgart, Deutschland

Einleitung

Die digitale Nachhaltigkeit von IT-Anwendungen in der Forschung spielt eine immer größer werdende Rolle, da IT-gestützte Forschungsergebnisse auch Jahre nach deren

Publikation reproduzierbar sein müssen, um Dritten das Nachvollziehen und Überprüfen der Ergebnisse zu ermöglichen. Wenn das Forschungsresultat auf der automatisierten Auswertung strukturiert dokumentierter Daten mittels Softwareprogrammen basiert, wird die stetige und zügige Weiterentwicklung von IT-Technologien jedoch zu einem immer größeren Problem: Werden Forschungsergebnisse beispielsweise mittels eines Windows 95-basierten Programms ermittelt, wird dessen Ausführung mit jeder neuen Generation von Betriebssystemen umständlicher, da sich Schnittstellen ändern und Annahmen nicht mehr erfüllt sind.

Während diese Probleme für einfache Softwareanwendungen mittels virtueller Maschinen gelöst werden können, sind komplexere Anwendungen mit diesem Ansatz nicht ohne großen manuellen Aufwand reproduzierbar. Basiert ein Forschungsergebnis beispielsweise auf einer umfangreichen softwarebasierten Simulation, welche unterschiedliche Dienste aufruft, die auf verschiedenen Betriebssystemen ausgeführt werden müssen, erfordert das Aufsetzen der Maschinen und Softwarekomponenten sowie deren Konfiguration detaillierte Expertise und ist mit großem Aufwand verbunden (Breitenbücher et al. 2013).

In diesem Beitrag zeigen wir auf, wie die standardbasierte open-source Technologie *OpenTOSCA* in den Digital Humanities eingesetzt werden kann, um die Reproduzierbarkeit IT-gestützter Forschungsergebnisse unabhängig von Technologieentwicklungen zu ermöglichen. Insbesondere verdeutlichen wir, wie auch komplexe Softwareanwendungen automatisiert bereitgestellt werden können, ohne detaillierte technische Expertise aufweisen zu müssen. Dadurch wird die nachhaltige Entwicklung von Forschungssoftware ermöglicht, indem diese auch Jahre später von Laien ausgeführt werden kann.

Nutzung des OpenTOSCA Ökosystems zur Sicherung der digitalen Nachhaltigkeit von Forschungsergebnissen

Das OpenTOSCA Ökosystem ist eine Werkzeugsammlung, welche die automatisierte Bereitstellung und Verwaltung von IT-Anwendungen ermöglicht. Die Werkzeuge basieren auf der *Topology and Orchestration Specification for Cloud Applications (TOSCA)* (OASIS 2013), einem OASIS Standard zur portablen Beschreibung von IT-Anwendungen. Der Standard definiert ein Metamodell zur Modellierung von *Anwendungsmodellen*, die alle Komponenten einer Anwendung, beispielsweise Webserver und Datenbanken, sowie deren Beziehungen untereinander beschreiben. TOSCA ist anbieter- und technologieagnostisch, wodurch ein Vendor-Lock-in verhindert wird. Dadurch können beliebige Komponententypen mittels TOSCA beschrieben und in Anwendungsmodellen miteinander kombiniert werden. Zur automatisierten Bereitstellung der modellierten Anwendungen definiert TOSCA die Konzepte der *Deployment Artifacts (DA)* und der *Implementation Artifacts*. Deployment Artifacts stellen die Implementierung einer Komponente dar. Beispielsweise kann die Java-Implementierung eines Analysealgorithmus als Deployment Artifact an das zugehörige Komponentenelement des Modells annotiert werden, siehe Abbildung 1. Managementoperationen, wie beispielsweise ein Installationsskript für einen Webserver, können mittels Implementation Artifacts modelliert werden. Um Anwendungsmodelle inklusive aller Artefakte zu paketieren, definiert TOSCA das selbstbeschreibende Archivformat *Cloud Service Archive (CSAR).*

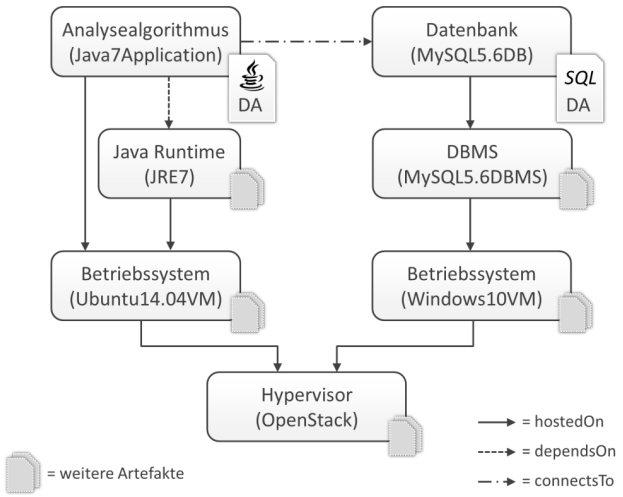


Abbildung 1: Simplifiziert dargestelltes Anwendungsmodell

Zur automatisierten Bereitstellung TOSCA-basierter Anwendungen werden TOSCA-Laufzeitumgebungen eingesetzt, welche die Anwendungsmodelle interpretieren und alle nötigen Bereitstellungsaktivitäten ausführen, d.h. modellierte virtuelle Maschinen

provisionieren, Webserver durch Ausführung von Implementation Artifacts installieren, Komponentenimplementierungen in Form von Deployment Artifacts ausliefern, etc. An der Universität Stuttgart wurde die open-source Laufzeitumgebung *OpenTOSCA* (Binz et al. 2013) sowie das TOSCA-Modellierungswerkzeug *Winery* (Kopp et al. 2013) entwickelt, um TOSCA-basierte Anwendungsmodelle auszuführen und zu erstellen. Das Selbstbedienungsportal *Vinothek* (Breitenbücher et al. 2014) ermöglicht es Nutzern, mittels eines Klicks, die Bereitstellung einer Anwendung zu veranlassen. Abbildung 2 zeigt das Zusammenspiel der Werkzeuge.

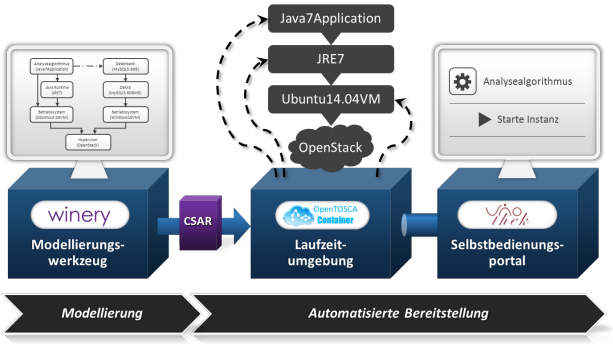


Abbildung 2: Werkzeuge des OpenTOSCA Ökosystems

Dieses OpenTOSCA Ökosystem kann zur Sicherung der digitalen Nachhaltigkeit von Forschungsergebnissen eingesetzt werden, indem Forschungssoftware in Form von CSARs paketiert wird. Durch die Möglichkeit, mit Winery alle erforderlichen Implementierungen in Form von Deployment und Implementation Artifacts zu spezifizieren, sowie die Struktur der Anwendung inklusive aller Beziehungen zwischen Komponenten zu modellieren, können Anwendungen selbstbeschreibend als CSAR archiviert werden. Diese CSARs können auch Jahre nach deren Entwicklung mittels der OpenTOSCA Laufzeitumgebung provisioniert werden, da alle nötigen Softwareartefakte und Modelle im CSAR enthalten sind und dadurch keine Abhängigkeiten zu externen Dateien existieren. Durch dieses Konzept können beispielsweise „Snapshots“ mehrerer virtueller Maschinen unterschiedlicher Betriebssysteme in Form von Virtual Machine Images in das CSAR gelegt und miteinander assoziiert werden, oder auch spezifische Webserver-Implementierungen, die Jahre später in der genutzten Form nur schwierig auffindbar sind bzw. von Laien nicht gemäß der erforderlichen Konfiguration installiert werden können. Die OpenTOSCA Laufzeitumgebung unterstützt zudem gängige Bereitstellungstechnologien wie Ansible (Hochstein 2014) oder Docker (Mouat 2015), wodurch Artefakte dieser Technologien ohne zusätzlichen Aufwand in das Anwendungsmodell eingebunden werden können. OpenTOSCA ermöglicht dadurch auch die effiziente Orchestrierung mehrerer Bereitstellungstechnologien.

Zur Reproduktion der Forschungsergebnisse muss die Software typischerweise mit auszuwertenden Forschungsdaten gestartet und parametrisiert werden. Häufig ist dies nicht trivial, beispielsweise wenn Data-Mining-Algorithmen auf Basis von Daten über Kostüme in Filmen wiederkehrende Muster finden sollen (Falkenthal et al. 2016). Das Konzept der CSARs ermöglicht auch diese Automatisierung, indem individuelle *Provisionierungspläne* für eine Anwendung modelliert werden können. Ein solcher Plan kann dann automatisiert von OpenTOSCA ausgeführt werden, um die Anwendung zu installieren und wie vorgesehen zu starten.

Bibliographie

Binz, Tobias / Breitenbücher, Uwe / Haupt, Florian / Kopp, Oliver / Leymann, Frank / Nowak, Alexander / Wagner, Sebastian (2013): „OpenTOSCA - A Runtime for TOSCA-based Cloud Applications“, in: *Proceedings of the 11th International Conference on Service-Oriented Computing (ICSOC 2013)*. Springer.

Breitenbücher, Uwe / Binz, Tobias / Kopp, Oliver / Leymann, Frank / Wettinger, Johannes (2013): „Integrated Cloud Application Provisioning: Interconnecting Service-Centric and Script-Centric Management Technologies“, in: *On the Move to Meaningful Internet Systems: OTM 2013 Conferences (CoopIS 2013)*. Springer.

Breitenbücher, Uwe / Binz, Tobias / Kopp, Oliver / Leymann, Frank (2014): „Vinothek - A Self-Service Portal for TOSCA“, in: *Proceedings of the 6th Central-European Workshop on Services and their Composition (ZEUS 2014)*. CEUR-WS.org.

Falkenthal, Michael / Barzen, Johanna / Breitenbücher, Uwe / Brügmann, Sascha / Joos, Daniel / Leymann, Frank / Wurster, Michael (2016): „Pattern Research in the Digital Humanities - How Data Mining Techniques Support the Identification of Costume Patterns“, in: *Proceedings of the 10th Symposium and Summer School On Service-Oriented Computing (SummerSOC)*. Springer.

Hochstein, Lorin (2014): *Ansible: Up and Running*. O'Reilly Media.

Kopp, Oliver / Binz, Tobias / Breitenbücher, Uwe / Leymann, Frank (2013): „Winery – A Modeling Tool for TOSCA-based Cloud Applications", in: *Proceedings of the 11th International Conference on Service-Oriented Computing (ICSOC 2013)*. Springer.

Mouat, Adrian (2015): *Using Docker: Developing and Deploying Software with Containers*. O'Reilly Media.

OASIS (2013): *Topology and Orchestration Specification for Cloud Applications Version 1.0*.

Digitale Werkzeuge und Infrastrukturen zur Analyse und Beschreibung von Bewegungen in vormodernen Wissensbeständen

Hegel, Philipp
hegel@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Tonne, Danah
danah.tonne@kit.edu
Karlsruher Institut für Technologie, Deutschland

Geukes, Albert
albert.geukes@cedis.fu-berlin.de
Freie Universität Berlin, Deutschland

Krewet, Michael
m.krewet@fu-berlin.de
Freie Universität Berlin, Deutschland

Rapp, Andrea
rapp@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Stotzka, Rainer
rainer.stotzka@kit.edu
Karlsruher Institut für Technologie, Deutschland

Uhlmann, Gyburg
g.uhlmann@fu-berlin.de
Freie Universität Berlin, Deutschland

Einführung

Der Sonderforschungsbereich 980 „Episteme in Bewegung" untersucht Prozesse des Wissenswandels in europäischen und nicht-europäischen Kulturen vom 3. Jahrtausend vor Christus bis ca. 1750 nach Christus. In den Analysen der insgesamt 22 Teilprojekte aus 20 Disziplinen wird gezeigt, wie gerade dort, wo in den Selbstbeschreibungen der vormodernen Kulturen und aus der Perspektive der Moderne Kontinuität und Stabilität im Vordergrund stehen, vielfältige Formen des Wandels und der Entwicklung beschrieben werden können. Wissen wird dabei als Episteme gefasst. Dieser Begriff aus der griechischen Antike schließt Wissenschaft ebenso ein wie nicht institutionalisierte Formen von Wissen. Er zeigt an, dass Wissen sich immer auf einen Gegenstand bezieht und impliziert zudem, dass Wissen als Wissen von etwas immer mit einem Geltungsanspruch versehen ist. Wissensbewegungen, die mit dem Terminus „Wissenstransfer" als Neukontextualisierung von Wissenselementen in neuen Kontexten beschreibbar gemacht werden, finden immer in komplexen Austauschprozessen statt, in denen verschiedene Akteure, Medien, Praktiken, Diskurse und Institutionen miteinander interagieren. Um diese komplexen multidirektionalen und multidimensionalen Prozesse erfassen und beschreiben zu können, werden digitale Werkzeuge entwickelt, die komplementär zu qualitativen exemplarischen Einzelanalysen auf größere Mengen an Texten und Bildern angewendet werden können.

Bücher auf Reisen

In diesem Rahmen entwickelt das Informationsinfrastrukturprojekt „Bücher auf Reisen" Softwarewerkzeuge, durch die räumliche und zeitliche Bewegungen von Handschriften, Drucken oder anderen Text- und Bildträgern auch für größere Objektzahlen hinweg systematisch erforscht werden. Die Ergebnisse werden als miteinander dynamisch vernetzte Elemente visualisiert. Auch „innere Reisen", das heißt Bewegungen in Objekten wie das Hinzufügen von Randnotizen oder der Verweis auf andere Texte, sollen auf diese

Weise digital aufbereitet und gespeichert werden. Neben dem Schwerpunkt auf der Entwicklung informationstechnisch unterstützter Verfahren zur Datenerschließung wird ein Forschungsdatenrepositorium für die digitalisierten Objekte mitsamt den neu ermittelten Metadaten zu Reisen und Veränderungsprozessen aufgebaut werden, das nachhaltig nutzbar ist.

Zentrale Fragestellungen sind hierbei die Modellierung und Verwaltung der Relationen der sehr heterogenen Datenbestände als „dynamische Metadaten“, eine verlässliche Speicherung aller Daten im Sinne einer Langzeitverfügbarkeit und eine sehr intuitive und benutzerfreundliche Bedienung. Durch die Verwendung international anerkannter Standards und Schnittstellen wird die Interoperabilität mit anderen standardisierten Infrastrukturen, zum Beispiel DARIAH-DE, und die leichte Erweiterbarkeit gewährleistet. Alle entwickelten oder adaptierten Softwarekomponenten werden der Öffentlichkeit als open source zur Verfügung gestellt.

Zuerst werden anhand von Pilotprojekten mit Daten aus der Aristotelesüberlieferung, altägyptischen Pyramidentexten, frühneuzeitlichen Fremdsprachenlehrwerken und einer Bibliothek des Osmanischen Reiches das Forschungsdatenrepositorium und die Softwarewerkzeuge erprobt und schrittweise verbessert, bis sie von allen Projektpartnertn verwendet werden können.

Handschriften in Bewegung

Im selben Rahmen werden im Gastprojekt „Handschriften in Bewegung“ digitale Verfahren zur Analyse von Veränderungen innerhalb von Handschriften im Sinne der genannten „inneren Reisen“ entwickelt. Mit Algorithmen der Bildverarbeitung aus dem Projekt „eCodicology“ können Merkmale des Layouts auf digitalisierten Buchseiten und Handschriften erkannt werden, die auch als strategische Momente zur Übermittlung von Wissen verstanden werden können. Auf der Grundlage dieser reproduzierbaren Messdaten werden durch statistische Auswertungen neue Erkenntnisse über Veränderungen in Einzelhandschriften oder Entwicklungen eines gesamten Buchbestandes gewonnen. Die Ergebnisse können insbesondere genutzt werden, um De- und Rekontextualisierungen von Wissensbeständen in Büchern aufzuzeigen.

Zu diesem Zweck werden verschiedene bereits existierende Softwarekomponenten eingesetzt:

Mit einem Bildverarbeitungsworkflow werden Seiten, Text- und Bildflächen auf Digitalisaten vermessen.

Mit einer Annotationssoftware können die vermessenen Bildbereiche geisteswissenschaftlich eingeordnet und mit zusätzlichen Informationen angereichert werden. So lässt sich die Wanderung einzelner Wissensbestände in einem Korpus nachvollziehen.

Die Metadaten werden graphisch aufbereitet, um gattungsspezifische Differenzen und historische Veränderungen sichtbar zu machen.

Fazit und Überlegungen zur Nachhaltigkeit

Die Forschungsfrage des Sonderforschungsbereichs „Episteme in Bewegung“ mit seiner Vielfalt an geisteswissenschaftlichen Disziplinen und Methoden wird durch den Einsatz informatischer Methoden und Fragestellungen substantiell bereichert. Umgekehrt bedeuten die Multidisziplinarität und die Diversität der Objekte ebenso eine Chance für die Weiterentwicklung informatischer Werkzeuge wie eine Herausforderung für die informatische Operationalisierung von Fragestellungen.

Auf Grund von Erfahrungen in früheren Projektarbeiten wird ein besonderer Fokus auf Nachhaltigkeit gelegt und diesem Bereich ein eigenes Arbeitspaket gewidmet. Im Rahmen einer engen Kooperation des SFB mit DARIAH-DE und dem Center für Digitale Systeme der Freien Universität Berlin stehen der nachhaltige Betrieb der erweiterten bzw. neu geschaffenen Forschungsdateninfrastruktur sowie deren fachliche Anschlussfähigkeit bzw. Anwendbarkeit auch für andere Fachrichtungen im Vordergrund. Durch standardisierte Schnittstellen sowie Nachnutzung und Erweiterung bestehender Werkzeuge werden bereits in der Konzeption erste Ansätze verfolgt, zusätzlich sind aber auch die Integration in bestehende Infrastrukturen und Institutionen vor Ort sowie in institutionenübergreifende Infrastrukturverbünde zentrale Fragestellungen.

Bibliographie

Chandna, Swati / Tonne, Danah / Jeikal, Thomas / Stotzka, Rainer / Krause, Celia / Vanscheidt, Philipp / Busch, Hannah / Prabhune, Ajinkya (2015): „Software workflow for the automatic tagging of medieval manuscript images (SWATI)“, in: Ringger, Eric K. / Lamiroy, Bart (eds.): *Proceedings SPIE9492, Document Recognition and Retrieval XXII, 940201* 10.1117/12.2076124 .

Chandna, Swati / Tonne, Danah / Stotzka, Rainer / Busch, Hannah / Vanscheidt, Philipp / Krause, Celia (2016): „An effective visualization technique for determining co-relations in high-dimensional medieval manuscripts data“, in: *Proceedings of Visualization and Data Analyses 2016* http://www.ingentaconnect.com/contentone/ist/ei/2016/00002016/00000001/art00013 .

Einfaches Topic Modeling in Python - Eine Programmbibliothek für Preprocessing, Modellierung und Analyse

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
Universität Würzburg, Deutschland

Schöch, Christof
christof.schoech@uni-wuerzburg.de
Universität Würzburg, Deutschland

Vitt, Thorsten
thorsten.vitt@uni-wuerzburg.de
Universität Würzburg, Deutschland

Topic Modeling ist eine Methode zur semantischen Erschließung größerer Textsammlungen, die in den letzten Jahren zunehmend in den Fokus der Aufmerksamkeit digital arbeitender Literaturwissenschaftler gerückt ist. Die Methode nutzt probabilistische Verfahren um aus einer Textsammlung eine Reihe von Verteilungen über die Wahrscheinlichkeiten einzelner Wörter zu erzeugen. Diese werden dann als distinkte semantische Gruppen, sogenannte ‘Topics’, aufgefasst, also als Gruppen inhaltlich zusammenhängender Wörter, die in den einzelnen Texten jeweils mehr oder weniger stark präsent sind (Blei 2012, Steyvers und Griffiths 2006).

Ursprünglich entwickelt, um in größeren Sammlungen kürzerer Fachartikel schnell jene zu identifizieren, die für bestimmte Themen relevant sein könnten, kann diese Methode darüber hinaus für eine Reihe von Problem im Bereich der digitalen Literaturwissenschaft interessante neue Lösungsansätze bieten. Dazu gehört die automatische Identifikation von Romanen, die ähnliche Themen behandeln (wenngleich eine direkte Gleichsetzung probabilistischer ‘Topics’ mit literarischen ‘Themen’ durchaus problematisch ist), ebenso wie die Zuordnung zu bestimmten Genres anhand inhaltlicher Aspekte, oder die quantifizierende Betrachtung der zu- und abnehmenden Bedeutung einzelner Themenfelder über den Verlauf eines einzelnen Romans (vgl. Blevins 2012, Jockers 2011, Rhody 2012, Schöch in Vorbereitung).

Mit den Programmen ‘Mallet’ (vgl. McCallum 2002) und ‘Gensim’ (vgl. Rehurek 2010) stehen zur Zeit zwei State-of-the-Art Implementierungen von Topic Modeling-Algorithmen zur Verfügung. Um die Methode produktiv einzusetzen, sind aber neben der Erzeugung des Modells weitere Arbeitsschritte notwendig (Abb. 1). Im ‘Preprocessing’ gilt es zunächst, die Textsammlungen in eine Form zu bringen, in der sie vom Modellierungsprogramm verarbeitet werden können. Darüber hinaus werden die Texte normalerweise durch das Herausfiltern häufiger Funktionswörter auf die potentiell inhaltsrelevanten Wörter reduziert, was in der Regel den vorhergehenden Einsatz von NLP-Tools (Natural Language Processing) erfordert. Sind die ‘Topics’ dann erst einmal errechnet worden, kann sich eine Visualisierung der Ergebnisse anschließen, oder ihre statistische Evaluierung anhand interner oder externer Kriterien, ein Aspekt dem beim Einsatz von Topic Modeling-Verfahren im DH-Kontext bisher eher zu wenig Beachtung geschenkt wurde.

Ziel unseres Projektes ist es, den Einstieg in aktuelle Topic Modeling-Verfahren für digital arbeitende Literaturwissenschaftler wesentlich zu vereinfachen, indem wir möglichst viele der notwendigen Arbeitsschritte in einer einheitlichen, umfangreichen und gut dokumentierten Programmbibliothek für die unter digital-quantitativ arbeitenden Geisteswissenschaftlern stark verbreitete Programmiersprache Python anbieten. Hierbei sollen Nutzerinnen und Nutzer bei allen Arbeitsschritten auf vorhandene, in einem ausführlichen Tutorial dokumentierte Funktionen zurückgreifen und so weit wie möglich wie mit einem Kommandozeilentool arbeiten können, ohne selbst programmieren zu müssen. Die Anforderungen an die Programmierkenntnisse der Forschenden, die diese Verfahren einsetzen möchten, werden damit minimiert und die Methode wird so einem größeren Nutzerkreis zugänglich gemacht.

Für das NLP-Preprocessing steht mit dem DARIAH-DKPro-Wrapper (DDW) ein komfortables Einheitswerkzeug zur Verfügung, das ein großes Spektrum an NLP-Aufgaben abdeckt und linguistische Annotationen in einem Python-Pandas-kompatiblen Ausgabeformat erzeugt. Ein Ziel unserer Bibliothek ist die direkte Anbindung des DDW-Outputs an existierende Implementierungen verschiedener etablierter Varianten von Topic Modeling-Algorithmen.

Für die Untersuchung der resultierenden Modelle möchten wir verschiedene Evaluierungsverfahren anbieten, sowohl interne Verfahren wie z.B. das Perplexity-Maß, als auch externe Vefahren, wie z.B. die Weglänge zwischen zwei Begriffen in einem Wörterbuch. Hieran schließen sich verschiedene Optionen zur Visualisierung der Ergebnisse an.

Im Fokus der Entwicklung steht die Gestaltung schlüssig aufeinander aufbauender Programmbefehle, die einer einheitlichen Syntax folgen und deren Funktion sich schnell erschließen lässt. Sie sollen sich ohne längere Einarbeitung nutzen und zu einer Pipeline zusammenfügen lassen, die die spezifischen Arbeitsschritte eines bestimmten Topic Modeling-Projektes umsetzt. Hierbei können Nutzerinnen und Nutzer auf detaillierte Anleitungen aus einem umfangreichen Tutorial zurückgreifen, in dem alle Funktionen, alle Outputs, und potentielle Kombinationen detailliert dokumentiert und anhand von Beispielen erläutert werden.

Die Entwicklung der Programmbibliothek kann auf Erfahrungen mit einer vorhandenen, Python-basierten Implementierung eines entsprechenden Workflows aufbauen, die allerdings eher “proof of concept”-Character hat (Topic Modeling Workflow “tmw”, vgl. Schöch 2015 und http://github.com/cligs/tmw).

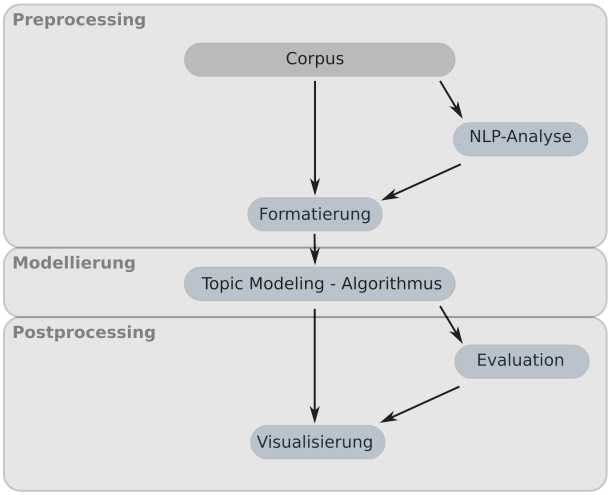


Abbildung 1: Workflow eines Topic Modeling-Projektes

Bibliographie

Blei, David M. (2012): „Probabilistic Topic Models“, in: *Communication of the ACM* 55 (4): 77–84 10.1145/2133806.2133826.

Blevins, Cameron (2010): „Topic Modeling Martha Ballard’s Diary“, in: *Historying* . http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/ .

Jockers, Matthew L. (2013): *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

McCallum, Andrew K. (2002): *MALLET: A Machine Learning for Language Toolkit* http://mallet.cs.umass.edu .

Rehurek, Radim / Sojka, Petr (2010): „Software framework for topic modelling with large corpora“, in: *Proceedings of LREC 2010*.

Rhody, Lisa M. (2012): „Topic Modeling and Figurative Language“, in: *Journal of Digital Humanities* 2 (1) http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/ .

Richardson, Stephen D. / Braden-Harder, Lisa (1988): „The Experience of Developing a Large-Scale Natural Language Text Processing System: CRITIQUE“, in: *Proceedings of the Second Conference on Applied Natural Language Processing* 195–202.

Schöch, Christof (in Vorbereitung): „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama“, in: *DHQ: Digital Humanities Quarterly* http://digitalhumanities.org/dhq . Preprint: https://zenodo.org/record/48356 .

Steyvers, Mark / Griffiths, Tom (2006): „Probabilistic Topic Models“, in: Landauer, T. / McNamara, D. / Dennis, S. / Kintsch, W.: *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.

Entitäten als Topic Labels: Verbesserung der Interpretierbarkeit und Evaluierbarkeit von Themen durch Kombinieren von Entity Linking und Topic Modeling

Lauscher, Anne
anne@informatik.uni-mannheim.de
Universität Mannheim, Deutschland

Nanni, Federico
federico@informatik.uni-mannheim.de
Universität Mannheim, Deutschland

Ponzetto, Simone Paolo
simone@informatik.uni-mannheim.de
Universität Mannheim, Deutschland

Im letzten Jahrzehnt haben Wissenschaftler aus dem Bereich der Geisteswissenschaften zunehmend mit verschiedenen Text Mining-Techniken zur Exploration großer Textkorpora experimentiert. Angefangen bei Kookkurrenz-basierten Verfahren (Buzydlowski, White und Lin 2002) über automatische Keyphrase Extraktion (Hasan, Saidul und Ng 2014) ziehen sich die angewandten Techniken bis hin zu Sequence Labeling Algorithmen, wie zum Beispiel im Falle von Named-Entity Recognition (Nadeau und Sekine 2007). Aus diesen vielfältigen Techniken bedienten sich die Forscher in den letzten Jahren vor allem des Latenten Dirichlet Allokation (LDA) Topic Model Algorithmus (Blei, Ng und Jordan 2003) (Meeks und Weingart 2012). Oftmals betonten Wissenschaftler dessen Potential für Serendipität (Alexander et al. 2014) und für Analysen im Bereich des Distant Reading (Leonard 2014; Graham, Milligan und Weingart 2016), also Studien, die über reine Textexploration hinausgehen.

In den letzten Jahren wurde LDA in den Digitalen Geisteswissenschaften intensiv angewandt, obwohl bekannt ist, dass die damit erzielten Ergebnisse schwierig zu interpretieren (Chang et al. 2009; Newman et al. 2010) und dass die Möglichkeiten, deren Qualität zu evaluieren, stark begrenzt sind (Wallach et al. 2009). Die direkte Konsequenz daraus ist, dass Wissenschaftler im Bereich der Geisteswissenschaften momentan in einer Situation feststecken, in der sie Topic Models weiterhin anwenden, da sie Methoden dieser Art benötigen, aber auch gleichzeitig nur wenig neues geisteswissenschaftliches Wissen ableiten können, weil die erzielten Ergebnisse bereits intrinsisch begrenzt sind (Nanni, Kümper und Ponzetto 2016). Diese Situation ist vor allem darauf zurückzuführen, dass große Korpora bestehend aus Primärquellen nun zum ersten Mal digital verfügbar sind.

Von dieser Grundsituation ausgehend wollen wir dieses komplexe Problem bewältigen, indem wir zwei spezifische und integrierte Lösungen zur Verfügung stellen. Als erstes bieten wir eine neue Methode zur Exploration von Textkorpora, die Topics erzeugt, welche leichter zu interpretieren sind als traditionelle LDA Topics. Dies erreichen wir durch die Kombination zweier Techniken, nämlich Entity Linking und Labeled LDA. Unsere Methode identifiziert in einer Ontologie eine Serie beschreibender Labels für jedes Dokument in einem Korpus. Daraufhin wird für jedes der identifizierten Labels ein Topic erzeugt. Durch die daraus resultierende direkte Beziehung zwischen Topic und Label wird die Interpretation des Topics stark vereinfacht und durch die Ontologie im Hintergrund wird die Ambiguität der Labels vermindert. Da unsere Topics mit einer limitierten Anzahl an klar umrissenen Labels beschrieben werden, fördern sie die Interpretierbarkeit und die Anwendung der Ergebnisse als quantitativ grundierte Argumente in der geisteswissenschaftlichen Forschung.

Da es äußerst wichtig ist, die Qualität der Ergebnisse zu bestimmen, stellen wir zweitens eine dreischrittige Evaluationsplattform zur Verfügung, die die Ergebnisse unseres Ansatzes

als Input verwendet und eine umfangreiche quantitative Analyse ermöglicht. Dies gestattet den nutzenden Wissenschaftlern aus den Digitalen Geisteswissenschaften, einen Überblick über die Ergebnisse der einzelnen Schritte der Pipeline zu erhalten und stellt Forschern im Natural Language Processing (NLP) eine Serie von Baselines zur Verfügung, die sie zur Verbesserung jedes Schrittes der vorgestellten Methodik benutzen können.

Wir illustrieren das Potenzial dieses Ansatzes durch dessen Anwendung zur Bestimmung der relevantesten Topics in drei verschiedenen Datensätzen. Der erste Datensatz besteht aus der gesamten Transkription der Reden aus dem fünften Mandat des Europäischen Parlaments (1999-2004). Dieses Korpus (van Aggelen et al. 2016) wurde für Forschung im Bereich der Computational Political Science bereits intensiv eingesetzt (Hoyland und Godbout 2008; Proksch und Slapin 2010; Høyland et al. 2014) und hat enormes Potential für zukünftige politikgeschichtliche Forschungen. Das zweite Korpus ist der sogenannte Enron-Datensatz. Es handelt sich dabei um eine große Datenbank mit über 600.000 E-Mails, die von 158 Mitarbeiten der Enron Corporation erstellt und die später durch die Federal Energy Regulatory Commission während der Untersuchungen nach dem Zusammenbruch des Unternehmens akquiriert wurden. In den letzten zehn Jahren hat die NLP-Community diesen Datensatz unter Anwendung von netzwerk- und inhaltsbasierten Analysen intensiv untersucht. Unser Ziel ist es hierbei, die Qualität unseres Ansatzes anhand eines hochtechnischen und komplexen Datensatzes einer spezifischen Art (E-Mail), die in zukünftigen historischen Untersuchungen immer wichtiger werden wird, zu beleuchten. In Verbindung damit wurde als drittes Korpus der Hillary Clinton E-Mail-Datensatz ausgewählt. Er repräsentiert eine Kombination der beiden vorherigen Datensätze, da es sich um kurze Korrespondenzen via E-Mail handelt, die sich jedoch mehrheitlich auf politische Themen fokussieren.

Vor über einem Jahrzehnt hat Dan Cohen (2006) bereits vorhergesehen, dass künftige Politikhistoriker in Anbetracht der Fülle an Quellen, die die öffentliche Verwaltung uns in den kommenden Jahrzehnten hinterlassen wird, auf ein Problem stoßen werden. Unsere Studie möchte ein allererster experimenteller Ansatz zu sein, diese neuen Korpora von Primärquellen zu bewältigen und Historiker im digitalen Zeitalter mit einer feinkörnigeren Lösung zur Textexploration als mittels traditionellen LDAs auszustatten.

Bibliographie

Alexander, Eric / Kohlmann, Joe / Valenza, Robin / Witmore, Michael / Gleicher, Michael (2014): „Serendip: Topic model-driven visual exploration of text corpora“, in: *IEEE VAST* 173–182.

Blei, David M / Ng, Andrew Y. / Jordan, Michael I. (2003): „Latent dirichlet allocation“, in: *Journal of Machine Learning Research* 3: 993–1022.

Buzydlowski, Jan W. / White, Howard D / Lin, Xia (2002): „Term co-occurrence analysis as an interface for digital libraries“, in: *Visual interfaces to digital libraries*. Springer 133–144.

Chang, Jonathan / Gerrish, Sean / Wang, Chong / Boyd-Graber, Jordan L. / Blei, David M. (2009): „Reading tea leaves: How humans interpret topic models“, in: *NIPS* 288–296.

Cohen, Dan (2006): *When machines are the audience*.

Graham, Shawn / Milligan, Ian / Scott Weingart (2016): *Exploring big historical data: The historian's macroscope*. Imperial College Press.

Hasan, Kazi Saidul / Ng, Vincent (2014): „Automatic keyphrase extraction: A survey of the state of the art“, in: *Proceedings of ACL-2014* 1262–1273.

Høyland, Bjørn / Godbout, Jean-François (2008): *Lost in translation? Predicting party group affiliation from European parliament debates*. Unveröff. Manuskript.

Høyland, Bjørn / Godbout, Jean-François / Lapponi, Emanuele / Velldal, Erik (2014): „Predicting party affiliations from European parliament debates“, in: *ACL 2014 Workshop on Language Technologies and Computational Social Science* 56–60.

Leonard, Peter (2014): „Mining large datasets for the humanities“ in: *IFLA WLIC* 16–22.

Meeks, Elijah / Weingart, Scott B. (2012): „The digital humanities contribution to topic modeling“, in: *Journal of Digital Humanities* 2 (1): 1–6.

Nadeau, David / Sekine, Satoshi (2007): „A survey of named entity recognition and classification“, in *Lingvisticae Investigationes* 30 (1): 3–26.

Nanni, Federico / Kümper, Hiram / Ponzetto, Simone Paolo (2016): „Semi-supervised textual analysis and historical research helping each other: Some thoughts and observations“ in: *International Journal of Humanities and Arts Computing* 10 (1): 63–77.

Newman, David / Lau, Jey Han / Grieser, Karl / Baldwin, Timothy (2010): „Automatic

evaluation of topic coherence“, in: *HLT-NAACL* 100–108.

Proksch, Sven-Oliver / Slapin, Jonathan B. (2010): „Position taking in European parliament speeches“, in: *British Journal of Political Science* 40 (3): 587–611.

van Aggelen, Astrid / Hollink, Laura / Kemman, Max / Kleppe, Martijn / Beunders, Henri (2016): „The debates of the European parliament as linked open data“, in: *Semantic Web* (Preprint) 1–10.

Wallach, Hanna M. / Murray, Iain / Salakhutdinov, Ruslan / Mimno, David (2009): „Evaluation methods for topic models“, in: *ICML* 1105–1112.

Grotefend digital

Vogeler, Georg
georg.vogeler@uni-graz.at
Karl-Franzens-Universität Graz, Österreich

Klugseder, Robert
robert.klugseder@oeaw.ac.at
Österreichische Akademie der Wissenschaften

Klug, Helmut W.
helmut.klug@uni-graz.at
Karl-Franzens-Universität Graz, Österreich

Steiner, Christian
christian.steiner@uni-graz.at
Karl-Franzens-Universität Graz, Österreich

Raunig, Elisabeth
elisabeth.raunig@uni-graz.at
Karl-Franzens-Universität Graz, Österreich

Die Entschlüsselung mittelalterlicher Datumsangaben setzt eine intensive Kenntnis des christlichen Oster- und Heiligenkalenders voraus. Historiker stützen sich bei der Berechnung von Zeitangaben in Handschriften und Urkunden, die in der Regel auf der Nennung von Kirchenfesten und deren Feier in bestimmten Regionen und Diözesen aufbauen, auf einschlägige Hilfsmittel, insbesondere den “Grotefend”. Zwischen 1891 und 1898 veröffentlicht der Historiker und Archivar Hermann Grotefend (1845-1931) sein Werk Zeitrechnung des deutschen Mittelalters und der Neuzeit in zwei Bänden, um damit sein veraltetes Handbuch der historischen Chronologie zu ersetzen. Es ist gleichzeitig die Quelle für das wiederholt aufgelegte auf die tägliche Praxis ausgerichtete Taschenbuch der Zeitrechnung des deutschen Mittelalters und der Neuzeit . Grotefends Monumentalwerk ist 2004 vom Archivar Horst Ruth retrodigitalisiert, für die Darstellung in HTML aufbereitet und – vor allem im Bereich des Heiligenverzeichnisses, das die Namen der Feste/Heiligen mit kalendarischen Daten und Ortsangaben verknüpft, – um eigene Forschungsergebnisse erweitert worden. Diese elektronische Ressource ist seit ihrer Entstehung ein beliebtes und einschlägig bekanntes Hilfsmittel. Darauf baut auch die semantische Modellierung des Grotefendschen Heiligenverzeichnisses auf: Die digitale Ressource des Grotefend wird durch semiautomatische Annotation (mithilfe von regulären Ausdrücken und XSLT) am Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities (ZIM-ACDH) als Semantic Web Resource (RDF) veröffentlicht.

Diese semantische Modellierung soll ermöglichen, Festkalenderangaben aus unterschiedlichen Kontexten eindeutig zu referenzieren. Damit kann eine verteilte Ressource aus den originalen Daten des Grotefend und RDF-Repräsentation von historisch belegten Kalendarien entstehen, die für kalenderbasierte Anwendungen nutzbar ist. Als Beispielanwendungen dienen eine digitale Kalenderedition (Teil der digitalen Edition des Tegernseer Wirtschaftsbuchs am Institut für Germanistik der Karl-Franzens-Universität Graz) und insbesondere das Projekt “Cantus Network - libri ordinarii of the Salzburg metropolitan province” der Österreichischen Akademie der Wissenschaften und des ZIM-ACDH. In letzterem werden mittelalterliche Libri ordinarii, deren liturgische Kalender bestimmten Regionen zugewiesen werden können, in XML modelliert und online publiziert. Durch den Abgleich mit dem Standardwerk zur historischen Datumsbestimmung können die Daten zu den einzelnen Heiligen, ihren Festtagen im Jahreskreis und ihrem jeweiligen regionalen Geltungsbereich durch die jeweiligen Handschriften der Libri ordinarii historisch verankert werden.

Ziel der Grotefend-Bearbeitung ist also, den digitalen Datenbestand nicht mehr nur als Text sondern auch als Forschungsdatenbank online zur Verfügung zustellen. Der Zugriff darauf soll über manuelle Suchabfragen von Heiligendaten ebenso wie über ein frei verfügbares Webservice als API möglich sein. Nach der semi-automatischen Modellierung des Grotefend ist eine Implementierung der

Ressource für den automatischen Vergleich von regional unterschiedlichen Kalendern vorgesehen. Bei einer derartigen Anwendung wird es möglich sein, die Beziehungen zwischen eindeutig referenzierbaren Heiligennamen, deren Festen (bzw. den dazugehörigen Kalenderdaten) und den davon betroffenen Regionen (Wirkungsbereiche, Diözesen, Orden) zu erkennen und grafisch darzustellen.

Das für die Konferenz geplante Poster legt den Fokus auf die Modellierung des Standardwerks in RDF und die daraus resultierenden Möglichkeiten für eine komplexe Suchoberfläche. Als Leitfragen stehen im Mittelpunkt: Wie weit können wenig strukturierte Textdaten möglichst automatisiert modelliert werden? Wie und mit welchem Aufwand sind komplexe, technisch unstrukturierte Daten in eine graphenbasierte Struktur zu überführen? Was sind die Voraussetzungen für die automatisierte Anreicherung von mittelalterlichen Kalendern?

Dabei werden aktuelle Methoden der digitalen Geisteswissenschaften angewandt: die Modellierung aller Daten in XML, die automatisierte Konvertierung der Kalenderdaten und des Grotefend in ein passendes RDF-Modell, die Zusammenführung dieser Datenquellen in einem Triplestore, die kontinuierliche Erweiterung der Datenbasis, die Abfrage der Daten mittels SPARQL.

Die damit geschaffene Ressource zur Bestimmung historischer Datumsangaben wird nach ihrer Fertigstellung der Öffentlichkeit frei zur Verfügung gestellt und soll durch die Modellierung in RDF Anreize bieten, weitere Kalendariendaten weltweit als Linked Open Data zur Verfügung zu stellen und somit zu einem großen gemeinsamen Datenbestand beitragen. Einen Anfang macht das Projekt, in dem es die Daten aus cantusdatabase.org, der Standardressouce zur Liturgieforschung, in RDF verwandelt und mit den Daten aus dem “großen Grotefend” verknüpft.

Bibliographie

Grotefend, Hermann (1970): *Zeitrechnung des deutschen Mittelalters und der Neuzeit*. 2 Bände. Neudruck: Aalen [Hannover 1891–1898].

Grotefend, Hermann (1891–1898): *Zeitrechnung des deutschen Mittelalters und der Neuzeit*. 2 Bände. 1891-1898. [Digitalisiert von Horst Ruth 2004]. http://bilder.manuscripta-mediaevalia.de/gaeste//Grotefend/kopf.htm [letzter Zugriff 16. August 2016].

W3C (2004): *Resource Description Framework (RDF)*. https://www.w3.org/RDF/ .

Ulrich, Theodor (1966): „Grotefend, Hermann“, in: *Neue Deutsche Biographie* 7. Berlin 165 f.

„IT for all“ – Das Projekt „Digitaler Campus Bayern – Digitale Datenanalyse in den Geisteswissenschaften“ als Beispiel für nachhaltige IT-Didaktik

Schulz, Julian
julian.schulz@lmu.de
Ludwig-Maximilians-Universität München, Deutschland

Ausgangslage

Der Einfluss der unter Digital Humanities (DH) zusammengefassten digitalen Theorien und Methoden auf die geisteswissenschaftlichen Disziplinen wächst stetig. Digitale Projekte erleben in den Geisteswissenschaften einen rasanten Aufschwung (Koller 2016: 43). Damit einher geht der Bedarf an Absolventen geisteswissenschaftlicher Fächer, die bereits während ihres Studiums Kompetenzen im Bereich der Digital Humanities erwerben konnten. Bereits 2013 forderten Vertreter/innen im „Manifest für die DH“ eine Etablierung „digitale[r] Trainingsprogramme in den Geisteswissenschaften“ (DH-Manifest: 2013), angepasst an die unterschiedlichen Bedürfnisse der Fachbereiche und die jeweiligen Karrierestufen. Auch der DHd misst der Ausgestaltung der IT-Ausbildung von Studierenden eine gesteigerte Bedeutung zu. Die Arbeitsgruppe zur Erarbeitung eines „Referenzcurriculums Digital Humanities“ [2] beschäftigt sich mit der Suche nach einer *bestpraxis*, von der Anwender und Institutionen gleichermaßen profitieren (Sahle 2013; Thaller 2015: 3).

Zahlreiche Universitätsstandorte haben auf die neuen Anforderungen mit der Einrichtung unterschiedlich ausgestalteter DH-Studiengänge

reagiert (Bartsch/Borek/Rapp 2016: 173; DH Course Registry). Trotz dieser neugeschaffenen Angebote besteht ein zusätzlicher Bedarf an informationstechnologischer Ausbildung in der Breite (Ehrlicher 2016: 625). Zunehmend wird auch in „klassischen“ geisteswissenschaftlichen Berufsfeldern Sicherheit im Umgang mit Software und digitalen Technologien vorausgesetzt. Dieses Grundverständnis digitaler Methoden kann nicht mehr ausschließlich im Selbststudium angeeignet werden (Spiro 2013: 332; Sahle 2016: 79).

Projektziele und Rahmenbedingungen

Hier setzt das Projekt „Digitaler Campus Bayern – Digitale Datenanalyse in den Geisteswissenschaften“ an, welches von der IT-Gruppe Geisteswissenschaften (ITG) der Ludwig-Maximilians-Universität München (LMU) seit Beginn dieses Jahres durchgeführt wird. Grundgedanke ist eine IT-Grundausbildung („*IT for all*“), welche die Studierenden problemorientiert in die Anwendung digitaler Methoden einführt. Ausgehend von fachwissenschaftlichen Fragestellungen werden Lehrveranstaltungen mit IT-Inhalten in Kooperation mit verschiedenen geschichtswissenschaftlichen Disziplinen und der Kunstgeschichte konzipiert. Dabei soll eine möglichst umfassend angelegte Grundlagenvermittlung in Erfassung, Modellierung, Analyse und anschließender Visualisierung von Daten erfolgen (Lücke/Riepl 2016: 77). Das Verständnis digitaler Methoden steht ebenso im Vordergrund wie eine fachliche Reflexion ihrer Potentiale (Rehbein 2016: 17).

Die Situation der DH an der LMU gestaltete sich bis Projektbeginn (Januar 2016 [3]) ambivalent. In den vorgenannten Studiengängen wurden regelmäßig Überblicksveranstaltungen zur Einführung in die Informatik für Historiker bzw. Kunsthistoriker angeboten. Eine praktische Umsetzung des theoretischen Wissens konnte im Rahmen dieser Veranstaltungen jedoch nicht geleistet werden. Demgegenüber werden durch die ITG, die auf langjährige und umfangreiche Erfahrungen im Bereich des digitalen Projektmanagements [4] verweisen kann, optimale Voraussetzungen für eine fortan praxisnahe IT-Ausbildung geschaffen.

Als Mitglied im Münchner Arbeitskreis für digitale Geisteswissenschaften (dhmuc) [5] kooperiert die IT-Gruppe zudem fach- und institutionsübergreifend mit zahlreichen kulturellen Einrichtungen. An der Schnittstelle zur universitären Lehre ist es möglich, die „*IT for all*“-Ausbildung geisteswissenschaftlicher Studierender auf die Anforderungen und Wünsche der potentiellen Arbeitgeberseite im (digitalen) Kultur-, Wissenschafts- und Informationssektor auszurichten.

Interaktive Lehr- und Lernumgebung *DHVLab*

Für die praktische Umsetzung kommt eine interaktive Lehr- und Lernumgebung, das *Digital Humanities Virtual Laboratory* – kurz *DHVLab* – zum Einsatz [6]. Die im Entstehen begriffene Plattform umfasst mehrere Komponenten, die im Folgenden vorgestellt werden sollen:

Virtuelle Rechenumgebung

Die virtuelle Rechenumgebung ist das „Herzstück“ der Ausbildungsplattform. Auf dem virtuellen Desktop werden in Abstimmung mit dem/der Kursleiter/in Software und Tools installiert. Dadurch wird die sukzessive Installation durch die Teilnehmer/innen obsolet, wodurch Probleme aufgrund unterschiedlicher Betriebssysteme und Versionierungen vermieden werden. Bei Anmeldung im *DHVLab* erhält jede/r Teilnehmer/in eine eigene SQL-Datenbank. Gleichzeitig werden strukturierteDatensammlungen vorgehalten. Diese sind für die Kursteilnehmer/innen zugänglich und für eigene oder im Kurs behandelte Fragestellungen verwendbar. Im Laufe der Lehrveranstaltung können neue Forschungsfragen ausgearbeitet und ein grundsätzliches Verständnis für den sinnvollen Einsatz von Tools und Software [7] in den Geisteswissenschaften entwickelt werden.

Ausbildungsmaterialien

Im vergangenen Semester wurde das System testweise in vorgenannten Einführungsveranstaltungen eingesetzt. Die bei der Evaluation gesammelten Erfahrungen fließen unmittelbar in die Erstellung bzw. Erweiterung der Ausbildungsmaterialien. Anhand praxisnaher Manuale wird IT-Grundlagenwissen, in einzelne Lehreinheiten gegliedert, anschaulich dargestellt und

erklärt. Die Erstellung von Lehrvideos und Übungsaufgaben ist vorgesehen. Aus diesem Portfolio können Dozentinnen und Dozenten Module entsprechend ihrer fachwissenschaftlichen Schwerpunktsetzung und der Voraussetzungen der Teilnehmer/innen auswählen. Die Seminarplanung und -durchführung erfolgt stets in enger Abstimmung mit den Projektmitarbeitern.

Publikationsumgebung

Für die Vor- und Nachbereitung der einzelnen Sitzungen steht ein WordPress-Blog zur Verfügung. Dort können die Kursleiter/innen Materialen einstellen, die Studierenden ihren Erkenntnisfortschritt und Analyseergebnisse dokumentieren. Dabei erlernen sie gleichzeitig *in praxi* das wissenschaftliche Bloggen als innovative Form des Publizierens. Eine abschließende Publikation der studentischen Seminararbeiten ist auf dieser Plattform möglich.

Datenrepositorium

In einem gesonderten Bereich der Datenbankumgebung werden die von den Studierenden im Rahmen einer Lehrveranstaltung erarbeiteten Datenbestände modelliert und nachhaltig abgelegt. Langfristiges Ziel ist der Aufbau eines Forschungsdatenrepositoriums. Nachfolgende Kurse mit ähnlichen Seminarthemen können auf diese Datensammlungen zugreifen, für die eigene Forschungsarbeit verwenden und dadurch sukzessive erweitern. Unterstützung erfährt die ITG durch die Universitätsbibliothek der LMU als Kooperationspartnerin auf dem Gebiet der nachhaltigen und nachnutzbaren elektronischen Publikation von Forschungsdaten.

Entwicklung eigener Analyse- und Softwarekomponenten

Mit dem *DHVLab Analytics Center* wurde eine Webanwendung entwickelt, die dazu dient, konkrete geisteswissenschaftliche Fragestellungen mithilfe quantitativer statistischer Methoden zu beantworten, sowie im Stile eines explorativen Werkzeuges neue Forschungsansätze zu eröffnen. Das *Analytics Center* kombiniert einführende deskriptive Analysen mit komplexeren Methoden der multivariaten Statistik. Neben diesem Analysetool entsteht derzeit eine Editionsumgebung, die speziell auf die Anforderungen von Studierenden und Promovierenden ausgerichtet wird. Diese wird erstmals im Sommersemester 2017 in einer Übung zur Edition mittelalterlicher Urkunden zum Einsatz kommen. Die Entwicklung weiterer Instrumente ist geplant.

Der Einsatz der Plattform in der Lehre

Nach der technischen Realisierung der Plattform und dem Aufbau grundlegender Ausbildungsmaterialien im ersten Projekthalbjahr kommt das System im Wintersemester 2016/2017 erstmals in eigens konzipierten Lehrveranstaltungen zur Anwendung. In der Kunstgeschichte soll das *Analytics Center* in einem Seminar zur Beschäftigung mit informatischen und mathematischen Verfahrensweisen anregen. Parallel dazu erfolgt eine Einführung in die Statistiksoftware *RStudio*. Ein geschichtswissenschaftliches Hauptseminar beschreitet den Weg von der Originalquelle über die strukturierte Aufnahme und Modellierung von Forschungsdaten sowie die Einführung in die Arbeit mit relationalen Datenbanken hin zur Georeferenzierung. Die in den Seminaren gewonnenen Erfahrungen und Erkenntnisse fließen unmittelbar in die Verbesserung und Ausweitung des bestehenden Lehrmaterials ein (u.a. Erstellung von Anwendungsszenarien). Neben den genannten Kursen wird die Plattform bereits in zahlreichen Lehrveranstaltungen als technische Grundlage verwendet [8].

Konzeption eines fachspezifischen DH-Curriculums

Die sukzessive wachsende Plattform und die aktuell angebotenen Kurse dienen als Grundlage für eine Institutionalisierung der IT-Grundausbildung in Form eines fachspezifischen DH-Curriculums. Das Konzept für das geplante freiwillige Zusatz-Zertifikat wird derzeit in der Projektgruppe erarbeitet und baut auf Erfahrungen vergleichbarer Angebote im deutschsprachigen Raum auf [9]. Angedacht

ist eine Kombination aus Veranstaltungen, die explizit IT-Grundlagenwissen vermitteln, und praxisorientierten Kursen, in denen die erlernten IT-Inhalte auf fachwissenschaftliche Gegenstände angewendet werden.
Wichtig erscheint eine ausgewogene Verschränkung von *eLearning*-Angeboten und Präsenzveranstaltungen, da insbesondere letzteren durch den intensiven Austausch der Studierenden mit DH-Spezialisten ein großer Beitrag zum Lernerfolg beigemessen wird [10] .

Grundlage einer nachhaltigen IT-Didaktik

Neben der Langzeitarchivierung der Forschungsdaten wird auch die Nachhaltigkeit der informationstechnologischen Infrastruktur (Serveranlage mit redundant ausgelegten File-, Datenbank- und Web-Servern sowie ausreichenden Storages) durch die IT-Gruppe Geisteswissenschaften dauerhaft gewährleistet. Die Architektur des *DHVLab* ist flexibel und skalierbar gestaltet, sodass sie weiter ausgebaut werden kann (bei Bedarf ist ein Hosting der Server am Leibniz-Rechenzentrum in Garching bei München möglich). Für eine nachhaltige IT-Didaktik spielt neben der langfristig gesicherten technischen Infrastruktur insbesondere die inhaltliche Kontinuität eine entscheidende Rolle. Die im Rahmen des Projektes erarbeiteten Lehreinheiten werden dauerhaft zur Verfügung gestellt. Thematisch sind sie so zu gliedern und fachlich anzupassen, dass eine spezifische Auswahl für eine Lehrveranstaltung und damit eine Integration in ein geisteswissenschaftliches Einzelfach möglich ist. Die IT-Gruppe stellt auch nach Ende der Projektlaufzeit die unterstützende Begleitung der Lehrveranstaltungen sicher.
Der Vortrag möchte zur Diskussion anregen, inwiefern sich die Anpassung der Materialien an die sich rasch wandelnden Anforderungen im Bereich der Digital Humanities möglichst effizient gestalten lässt. IT-Didaktik scheint nur dann einen Anspruch auf Nachhaltigkeit zu besitzen, wenn sie sich in einem steten Anpassungsprozess befindet.

Ganz im Sinne des „Digitalen Campus Bayern“ ist das Münchener Pilotprojekt auf eine Ausweitung auf andere Studienstandorte ausgerichtet. Die Plattform wird beispielsweise ab 2017 in einem im Aufbau befindlichen Kooperationsprogramm zur DH-Ausbildung der Universitäten Erlangen, München und Regensburg zum Einsatz kommen. Alle Module des *DHVLab* können kollaborativ von anderen Hochschulen genutzt werden, um umfassende Sammlungen von Tutorials, Aufgaben, Softwarebeschreibungen, Anwendungsszenarien sowie Sammlungen fachwissenschaftlicher Objekt- und Metadaten aufzubauen und gemeinsam zu pflegen.

Fußnoten

2. Vgl. http://www.dh-curricula.org/index.php?id=1 [letzter Zugriff: 30. November 2016].
3. Die Projektlaufzeit beträgt zwei Jahre. Das Vorhaben ist Teil eines Förderprogramms, welches das Bayerische Wissenschaftsministerium aufgelegt hat. Vgl. https://www.km.bayern.de/pressemitteilung/9340/.html [letzter Zugriff: 30. November 2016].
4. Vgl. die Übersicht unter www.itg.lmu.de/projekte [letzter Zugriff: 30. November 2016].
5. Vgl. http://dhmuc.hypotheses.org/uber [letzter Zugriff: 30. November 2016].
6. Für die Dokumentation der technischen Infrastruktur vgl. http://dhvlab.gwi.uni-muenchen.de/index.php/Category: Architektur [letzter Zugriff: 30. November 2016].
7. Derzeit stehen in der virtuellen Umgebung u.a. folgende Software und Programme zur Verfügung: LibreOffice-Paket, OCRFeeder und Ocrad (Texterkennung), Python (PyCharm), RStudio (Statistik), Gephi (Visualisierung), epcEdit (XML-Editor), AntConc und TreeTagger (Korpuslinguistik).
8. Vgl. die Zusammenstellung auf der Projektseite: http://dhvlab.gwi.uni-muenchen.de/index.php/Das_DHVLab_im_Einsatz [letzter Zugriff: 30. November 2016].
9. Vgl. insbesondere die Angebote in Köln (http://www.itzertifikat.uni-koeln.de/), Passau (http://www.phil.uni-passau.de/zertifikat-dh/) und Stuttgart („Das digitale Archiv“, http://www.uni-stuttgart.de/dda), letztgenanntes als Vorläufer eines DH-Masterstudienganges [letzter Zugriff: 30. November 2016].
10. Vor diesem Hintergrund erscheinen grundständige *eLearning*-Angebote wie „The Programming Historian“ (http://programminghistorian.org/) für einen autodidaktischen Einstieg begrüßenswert. Die Initiatoren des DHVLab sind jedoch der Auffassung, dass eine umfassende Präsenzausbildung nicht ersetzt werden kann.

Bibliographie

Bartsch, Sabine / Borek, Luise / Rapp, Andrea (2016): „Aus der Mitte der Fächer, in die Mitte der Fächer: Studiengänge und Curricula – Digital Humanities in der universitären Lehre“, in: *Bibliothek – Forschung und Praxis* 40 (2): 172–178 10.1515/bfp-2016-0030.

DARIAH-EU: *Digital Humanities Registry – Courses* https://dh-registry.de.dariah.eu/ [letzter Zugriff 30. November 2016].

DHI Paris (Teamaccount) (2013): „Wissenschaftlicher Nachwuchs in den Digital Humanities: Ein Manifest“, in: *Digital Humanities am DHIP*, 23. August 2013 http://dhdhi.hypotheses.org/1995 [letzter Zugriff 30. November 2016].

Ehrlicher, Hanno (2016): „Fingerübungen in Digitalien. Erfahrungsbericht eines teilnehmenden Beobachters der Digital Humanities aus Anlass eines Lehrexperiments“, in: *Romanische Studien* 4: 623–636 http://www.romanischestudien.de/index.php/rst/article/view/88 [letzter Zugriff 30. November 2016].

Koller, Guido (2016): *Geschichte digital: Historische Welten neu vermessen*. Stuttgart: Kohlhammer.

Lücke, Stephan / Riepl, Christian (2016): „Auf dem Weg zu einem Curriculum in den Digital Humanities“, in: *Akademie Aktuell* 57 (1): 74–77 http://badw.de/fileadmin/pub/akademieAktuell/2016/56/0116_17_Riepl_V04.pdf [letzter Zugriff 30. November 2016].

Rehbein, Malte (2016): *Geschichtsforschung im digitalen Raum. Über die Notwendigkeit der Digital Humanities als historische Grundwissenschaft*. (Preprint) http://www.phil.uni-passau.de/fileadmin/dokumente/lehrstuehle/rehbein/Dokumente/GeschichtsforschungImDigitalenRaum_preprint.pdf [letzter Zugriff 30. November 2016].

Sahle, Patrick (2013): *DH studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities (= DARIAH-DE Working Papers 1)*. Göttingen: GOEDOC http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2013-1.pdf [letzter Zugriff 30. November 2016].

Sahle, Patrick (2016): „Digital Humanities als Beruf. Wie wird man ein „Digital Humanist“, und was macht man dann eigentlich?“, in: *Akademie Aktuell* 57 (1): 78–83 http://badw.de/fileadmin/pub/akademieAktuell/2016/56/0116_18_Sahle_V04.pdf [letzter Zugriff 30. November 2016].

Spiro, Lisa (2012): „Openingup Digital Humanities Education“, in: Hirsch, Brett D. (ed.): *Digital Humanities Pedagogy: Practices, Principlesand Politics* 331–363 http://www.openbookpublishers.com/product/161/ [letzter Zugriff 30. November 2016].

Thaller, Manfred (2015): „Panel: Digital Humanities als Beruf – Fortschritte auf dem Weg zu einem Curriculum“, in: *Digital Humanities als Beruf: Fortschritte auf dem Weg zu einem Curriculum, vorgelegt auf der Jahrestagung 2015* 3–5 https://www.digitalhumanities.tu-darmstadt.de/fileadmin/dhdarmstadt/materials/Digital_Humanities_als_Beruf_-_Stand_2015.pdf [letzter Zugriff 30. November 2016].

Kollaborative Forschung über Linked Open Data Forschungsdatenbanken der Universitätsgeschichte Implementierung des Heloise Common Research Model

Riechert, Thomas
thomas.riechert@htwk-leipzig.de
Hochschule für Technik, Wirtschaft und Kultur (HTWK) Leipzig, Deutschland

Beretta, Francesco
francesco.beretta@ish-lyon.cnrs.fr
Laboratoire de recherche historique Rhône-Alpes (LARHRA), Frankreich

Motivation

Die Beantwortung von Forschungsfragen über bestehende Forschungsdatenquellen im Linked Open Data Web ist von besonders hohe Relevanz für nachhaltige Forschungsarbeiten in den Geisteswissenschaften und insbesondere in der Geschichtswissenschaft (vgl. Meroño Peñuela et al. 2014: 1-27). Eine ständig wachsende Menge an Archivmaterialien, Literaturquellen, Forschungsergebnisse und Forschungsdatenbanken ist online verfügbar.

Durch Standardisierungsbestrebungen u. a. durch den Einsatz von RDF und OWL als Beschreibungssprache von Daten ist es möglich diese zu verknüpfen, und Inferenz-Algorithmen auf diesen Ressourcen anzuwenden. Eine typische Herangehensweise an die damit einhergehenden Herausforderungen der Datenintegration ist die Verwendung von Standard-Vokabularen, wie sie z. B. im E-Business erfolgreich praktiziert wird (vgl. Domingue et al. 2011). Im Bereich der Geisteswissenschaften haben sich zum Beispiel mit GND [1] und Europeana Data Model (EDM) [2] solche Vokabulare entwickelt. Darüber hinaus wird CIDOC-CRM [3] von unterschiedlichen Projekten als Modell für die Veröffentlichung von Daten verwendet (vgl. Kurtz et al. 2009). Die Autoren sind in verschiedenen Forschungsprojekten in Deutschland und Frankreich im Bereich der Informatik und der Geschichtswissenschaften aktiv und müssen konstatieren, dass der Weg der Standardisierung in der historischen Forschung schwieriger ist. Dies ist vor allem auf den hohen Grad der domänspezifischen Eigenheiten der vorliegenden Daten, und auf die besondere Rolle projektbezogener Forschungsfragestellungen bei der Datenerstellung und Datenerhebung zurückzuführen (vgl. Beretta 2009).

Das im Jahr 2012 gegründete Europäische Netzwerk Héloïse [4] zur Vernetzung von online verfügbaren Datenbanken im Bereich der Universitätsgeschichte [5], sieht sich dieser Herausforderung gegenübergestellt. Inhalte regelmäßig stattfindender Workshops sind die Präsentation von verfügbaren Forschungsdatenbanken und deren Verwendung bei auf diese heterogenen Datenbanken aufbauenden Forschungsfragestellungen. Parallel mit der Nutzung der Repositorien in zukünftigen Forschungskontexten erfolgt einekollaborative Daten-Kuration und sichert daher die Langzeitlebigkeit von historischen Forschungsdaten - nicht nur im Sinne von Datensicherung, sondern auch von ständiger Anreicherung und Qualitätsverbesserung der Daten.

Heloise Common Research Model

Mit dem Heloise Common Research Model (HCRM) haben die beiden Autoren dem Konsortium eine Methode präsentiert (vgl. Beretta und Riechert 2016), welches durch die Forschungspartner gemeinsam verfeinert wird und in der Forschungsdomäne eingesetzt werden soll. Das HCRM ist als Schichtenarchitektur konzipiert, bestehend aus drei Schichten: dem Repository-Layer, dem Application-Layer und dem Research-Interface-Layer. In einem parallelem Entwicklungsprozess entsteht neben der detaillierten Definition von Modulen (vgl. Abbildung 1, Beretta und Riechert 2016) im HCRM eine Implementierung der Methodologie in Form der Héloïse-Plattform.

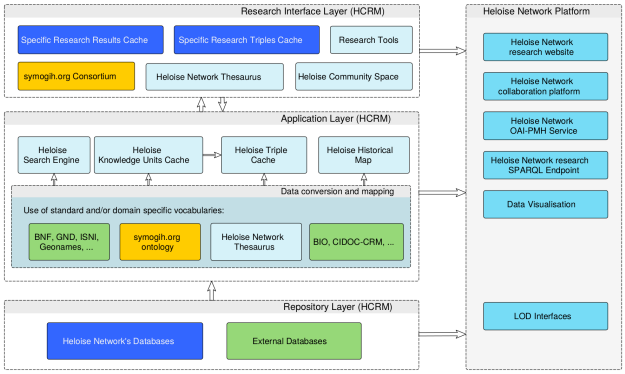


Abbildung 1: Heloise Common Research Modell – Überblick über die Schichten und Module (links), Héloïse-Plattform (rechts) (vgl. Beretta und Riechert 2016)

Repository-Layer Die Datenbanken der Partnern des Héloïse-Netzwerkes, genauso wie externe Informationsquellen, stellen die Basis zur Beantwortung übergeordneter Forschungsfragen. Die Publikation der Daten als LOD erfolgt mit Hilfe etablierter Werkzeuge. So werden u. a. die Werkzeuge OntoWiki (Frischmuth et al. 2013) und D2RQ (Bizer und Cyganiak 2006) zum Publizieren eingesetzt. Die Verlinkung erfolgt innerhalb etablierter Authoring-Prozesse in den Forschungsprojekten.

Application-Layer: Die Applikations schicht unterstützt das Auffinden von Ressourcen und deren Verlinkung. Hierfür wird auf generische Werkzeuge und standardisierte Vokabulare aufgebaut, damit einhergehend ist der Zugang auf die Informationen auf diese Standards beschränkt. Als erste Anwendungen im Forschungskonsortium wird gegenwärtig die Implementierung einer Personen-Suche, basierend auf der BIO-Ontologie (Davis und Galbraith 2010), sowie die Darstellung relevanter geographischer Daten, über die Repositorien der Partner, entwickelt.

Research Interface Layer Die Forschungsschicht des Modells bietet einen Zugang auf die Forschungsdaten für neue

wissenschaftliche Fragestellungen. Hierfür wird basierend auf einer Meta-Ontologie, wie sie das SyMoGIH Projekt (Beretta 2015, Beretta 2016) für die Geschichtswissenschaft entwickelt hat [6], ein fachspezifisches Meta-Vokabular entwickelt. Dieses Vokabular verbindet einen kollaborativen, domänespezifischen Ansatz mit einer größtmöglichen Unabhängigkeit von spezifischen, Projekt-bezogenen Forschungsfragestellungen und ermöglicht den Datenaustausch im Héloïse-Netzwerk.

Héloïse-Plattform

Zentral für zukünftige Forschungsprojekte im Héloïse-Konsortium ist die Implementierung der Methodologie. Die heterogene Zusammensetzung des Konsortium spricht gegen die Etablierung einer zentralen Administration einer solchen Plattform. Vielmehr ist die Forschungskooperation auf der Ebene der Historiker in einer vergleichbaren Art und Weise bei den assoziierten IT-Partnern zu finden. Die Autoren schlagen daher die Realisierung von Diensten der Plattform durch Microservices vor. Microservice stellen unabhängige Dienste zur Verfügung und sind im Sinne der angestrebten Cloud-basierten Plattform virtualisierbar (vgl. Newman 2015).

Das Poster stellt die Resultate der iterativen Implementierung des HCRM im Héloïse-Forschungsnetzwerk detailliert vor. Es werden die Ergebnisse der fachlichen Diskussion über zwei Heloise Workshops (Madrid, 2015 und Perugia, 2016) vorgestellt. Der komplette Katalog, der durchgängig im Kontext der Linked-Open-Data-Philosophie, als Open-Source verfügbaren Microservices (Docker Container [7]), wird präsentiert und deren Anwendung innerhalb der Heloise-Plattform online gezeigt [8].

Fußnoten

1. GND Ontology http://d-nb.info/standards/elementset/gnd [25/08/2016]
2. Europeana Data Model: http://pro.europeana.eu/page/edm-documentation [25/08/2016]
3. CIDOC Conceptual Reference Model: http://www.cidoc-crm.org [25/08/2016]
4. Héloïse - European Workshop on academic Database: http://heloisenetwork.eu/ [25/08/2016]
5. Héloïse-Partner: http://heloisenetwork.eu/repositories [25/08/2016]
6. Ontologie und Instanzen des Vokabulars sind auf der Webseite des Projektes online zugänglich: http://symogih.org [25/08/2016]
7. Docker Virtualisierung: http://docker.com [25/08/2016]
8. Héloïse-Network-Plattform: http://heloisenetwork.eu/platform [25/08/2016]

Bibliographie

Beretta, Francesco (2015): „Publishing and sharing historical data on the semantic web: the SyMoGIH project–symogih.org“, in: *Workshop: Semantic Web Applications in the Humanities* II.

Beretta, Francesco (2016): „L'interopérabilité des données historiques et la question du modèle: l'ontologie du projet SyMoGIH“, in: Minel, Jean-Luc (ed.): *Quels enjeux numériques pour les médiations scientifique et culturelle*. Presses universitaires de Paris Ouest (im Erscheinen).

Beretta, Francesco / **Riechert, Thomas** (2016): „Collaborative Research on Academic History using Linked Open Data: A Proposal for the Heloise Common Research Model“, in: *CIAN-Revista de Historia de las Universidades, Norteamérica* 19 (Juni).

Bizer, Christian / **Cyganiak, Richard** (2006): „D2r server-publishing relational databases on the semantic web“, in: *5th International Semantic Web Conference* 294–309.

Davis, Ian / **Galbraith, David** (2010): *BIO: A vocabulary for biographical information*.

Domingue, John / **Fensel, Dieter** / **Hendler, James A.** (eds.) (2011): *Handbook of semantic web technologies*. Berlin / Heidelberg: Springer Science / Business Media.

Frischmuth, Philipp / **Martin, Michael** / **Tramp, Sebastian** / **Riechert, Thomas** / **Auer, Sören** (2014): „OntoWiki - An Authoring, Publication and Visualization Interface for the Data Web“, in: *Semantic Web Journal (IOS Press)*.

Kurtz, Donna / **Parker, Greg** / **Shotton, David** / **Klyne, Graham** / **Schroff, Florian** / **Zisserman, Andrew** / **Wilks, Yorick** (2009): „Claros-bringing classical art to a global public“, in: *Fifth IEEE International Conference on e-Science* 20–27.

Meroño-Peñuela, Albert / **Ashkpour, Ashkan** / **van Erp, Marieke** / **Mandemakers, Kees** / **Breure, Leen** (2014): „Semantic technologies for historical research: A survey“, in: *Semantic Web Journal (IOS Press)*.

Newman, Sam (2015): *Building Microservices: Designing Fine-Grained Systems*. O'Reilly Media.

Kompilation eines Diskursstruktur-annotierten deutschsprachigen Blogkorpus

Grumt Suárez, Holger
Holger.H.Grumt-Suarez@germanistik.uni-giessen.de
Justus-Liebig-Universität Gießen, Deutschland

Karlova-Bourbonus, Natali
Natali.Karlova-Bourbonus@germanistik.uni-giessen.de
Justus-Liebig-Universität Gießen, Deutschland

Lobin, Henning
Henning.Lobin@uni-giessen.de
Justus-Liebig-Universität Gießen, Deutschland

Das Poster „Interoperabilität bei der Erstellung eines deutschsprachigen Blogkorpus für die Repräsentation der Diskursstruktur“ informiert über das Vorgehen sowie die ersten Forschungszwischenergebnisse und die weiteren Ziele der Kompilierung und Annotation eines deutschsprachigen Blogkorpus. Gegenwärtig gibt es lediglich eine geringe Anzahl an öffentlich zugänglichen, umfangreichen Blogkorpora, wie zum Beispiel das englischsprachige Birmingham Blog Corpus der Birmingham Universität (vgl. WebCorp 2013) oder das bilinguale (deutsch-französische) Korpus d’apprentissage INFRAL (Interculturel Franco-Allemand en Ligne) (vgl. Abendroth-Timmer et al. 2014). Betrachtet man den großen Einfluss von Blogs für die Geschichte der Kommunikation im Internet, erscheint die geringe Anzahl überraschend.

Bislang existiert kein Standard für das Repräsentieren von sogenannten Computer-Mediated Communication-Daten (kurz CMC), allerdings arbeitet die Text Encoding Initiative CMC Special Interest Group (TEI SIG) (vgl. Beißwenger 2016) seit 2013 an einem Schema für die Repräsentation von CMC-Genres. Eine Standardisierung, wie sie die Text Encoding Initiative für CMC anstrebt, ist ein wichtiger Punkt, wenn es um ‚Digitale Nachhaltigkeit‘ geht. Unser Forschungsvorhaben leistet hierfür einen Beitrag.

Das Hauptziel des Vorhabens umfasst die semi-automatische Kompilation sowie die Repräsentation der Blogdiskursstruktur. Dabei sollen die Relationen zwischen den textuellen und multimodalen Elementen (Blogbeiträge, Kommentare, Hyperlinks, Bilder und Töne) und den verschiedenen Textproduzenten (Blogger, Kommentatoren) abgebildet werden. Das annotierte Blogkorpus soll am Ende als eine nachhaltige Ressource verfügbar gemacht werden. Hierfür stehen wir momentan in Kontakt mit der Redaktion von Spektrum der Wissenschaft Verlagsgesellschaft mbH, um die Form der Bereitstellung zu klären.

Die Grundlage des Korpus bildet das Wissenschaftsblogportal SciLogs – Tagebücher der Wissenschaft (SciLogs 2016) und deckt den Inhalt des Jahres 2015 vollständig ab. Die Daten wurden aus den vier SciLogs-Blogbereichen „WissensLogs“, „BrainLogs“, „KosmoLogs“ und „ChronoLogs“ erhoben. Das Korpus soll mit drei Informationstypen annotiert werden, die einerseits direkt, andererseits indirekt in den Blogdaten vorhanden sind oder anhand von statistischen Analysen und computerlinguistischen Tools sichtbar gemacht werden. Zum jetzigen Zeitpunkt beschränken sich die Annotationen des Korpus auf die direkt auslesbaren Informationen des Blogs wie beispielsweise der Titel des Blogbeitrags, der Name des Bloggers und das Einstelldatum des Blogeintrags. Ferner wird darauf geachtet, dass sämtliche Informationen, die die Inhalte der Scilogs-Website in Bezug auf die Bloginhalte liefern, ebenfalls annotiert werden. Wir sind der Meinung, dass auch auf den ersten Blick nicht für die Diskursstruktur relevante Informationen ausgezeichnet werden sollten. Im Fokus steht der Ansatz, dass das Blogkorpus aus CMC-Daten später für die Erforschung unterschiedlicher linguistischer Fragestellungen verwendet werden kann.

Zusammenfassend soll das Poster nicht nur unser Vorhaben vorstellen, sondern auch einen Einblick in unser grundsätzliches Vorgehen bei der Erstellung eines CMC-Korpus geben. Der Fokus für die DHd 2017 liegt unter anderem auf der Darstellung der Entscheidungsfindung innerhalb der Auszeichnungssprachen. Es soll erläutert werden, warum wir uns beispielsweise für die deskriptive Auszeichnungssprache TEI und nicht XML (Extensible Markup Language) entschieden haben. Des Weiteren möchten wir Einblicke in die semi-automatische TEI-Annotation geben und unsere Erkenntnisse mit dem vorläufigen, von der TEI SIG bereitgestellten, TEI-Schema teilen. Letztlich wollen wir auch das bisherige

Korpus selbst vorstellen, das aus ca. 3.000.000 Tokens (ca. 1.200 Blogposts von 80 Bloggern und 15.000 Kommentaren von 1500 Kommentatoren) besteht.

Bibliographie

Abendroth-Timmer, Dagmar / Bechtel, Mark / Chanier Thierry / Ciekanski, Maud (2014): *Corpus d'apprentissage INFRAL (Interculturel Franco-Allemand en Ligne)*. Banque de corpus CoMeRe. Nancy: Ortolang.fr https://hdl.handle.net/11403/comere/cmr-infral [letzter Zugriff 1. Juli 2016].

Beißwenger, Michael (2016): *SIG: Computer-Mediated Communication* http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication [letzter Zugriff 1. Juli 2016].

SciLogs (2016): *SciLogs: Tagebücher der Wissenschaft*. Spektrum der Wissenschaft Verlagsgesellschaft mbH http://www.scilogs.de/impressum/ [letzter Zugriff 1. Juli 2016].

WebCorp (2013): *Birmingham Blog Corpus. WebCorp: Linguist's Search Engine*. Birmingham City University http://wse1.webcorp.org.uk/cgi-bin/BLOG/index.cgi [letzter Zugriff 1. Juli 2016].

Kriterienbasierte Evaluation und Dokumentation technischer Nachhaltigkeit von Forschungssoftware in einem Metadatenrepositorium

Druskat, Stephan
stephan.druskat@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Einleitung

Softwarenachhaltigkeit kann im Bezug auf verschiedene Aspekte definiert werden: Umwelt, Gesellschaft, Wirtschaft, Technik, Individuen (vgl. Becker et al. 2015). Von besonderem Interesse für die Forschung ist dabei die technische Nachhaltigkeit von Forschungssoftware, d.h., ob für die eigene Forschungsfrage verwendbare Software existiert, ob diese eigenen Bedürfnissen anpassbar ist, inwiefern ihre Nachnutzbarkeit - und damit auch die Reproduzierbarkeit von Forschungsergebnissen - in Zukunft gesichert ist, etc. Praktisch spielen dabei vor allem zwei Kriterien eine Rolle: die Sichtbarkeit nachnutzbarer Forschungssoftware und die nachvollziehbare Bewertung ihrer technischen Nachhaltigkeit.

Parallel zur Dokumentation der Nachhaltigkeit von Forschungsdaten in Datenmanagementplänen bieten sich Softwaremanagementpläne zur Dokumentation der technischen Nachhaltigkeit neu erstellter Forschungssoftware an und es ist anzunehmen, dass Förderer in Zukunft diesen Weg einschlagen werden (vgl. Hettrick 2015). Zur Dokumentation von Forschungssoftware bieten sich - ebenfalls parallel zu Forschungsdaten - Repositorien an, die mit Hilfe von Metadaten Forschungssoftware sichtbar und für eine Verwendung bewertbar machen.

State of the art

Solche Metadatenrepositorien existieren bereits, im geisteswissenschaftlichen Bereich beispielsweise DiRT Directory und CLARIN Virtual Language Observatory, für andere oder mehrere Disziplinen SciencePAD und EGI Applications Database (vgl. auch re3data.org). Softwaremetadaten sind ebenfalls zugänglich über Plattformen wie GitHub, Open Hub und Zenodo. Während diese Werkzeuge für die Suche nach verwendbarer Foschungssoftware prinzipiell, wenn auch teilweise nur eingeschränkt, geeignet sind, dokumentiert keins von ihnen umfassend den Grad der technischen Nachhaltigkeit der präsentierten Software ohne die Notwendigkeit massiver Interpretation und Extrapolation seitens der Nutzerin.

Ein messendes Metadatenrepositorium für Forschungssoftware

Um sowohl die Sichtbarkeit als auch die Einschätzbarkeit der technischen Nachhaltigkeit von Forschungssoftware zu gewährleisten

bietet sich daher die Konzeption eines Typus von Metadatenrepositorium an, der ausgehend von Druskat (2016) hier beschrieben und weiterentwickelt wird. In einem solchen Repositorium werden hinterlegte Metadaten unter Zuhilfenahme von Nachhaltigkeitskriterien quantifiziert und auf dieser Grundlage Maße berechnet, die nachvollziehbar und reproduzierbar den Grad technischer Nachhaltigkeit abbilden können. Dabei muss ein Metadatenbegriff zu Grunde gelegt werden, der weit genug ist, relevante Informationen für die Bestimmung von sowohl (Nach-)Nutzbarkeit als auch technischer Nachhaltigkeit erfassen zu können.

Voraussetzungen

Für die Umsetzung eines solchen Repositoriums müssen mindestens vier zentrale theoretische Probleme gelöst werden, nämlich 1. wie technische Nachhaltigkeit operationalisierbar definiert werden kann, 2. wie Kriterien für technische Nachhaltigkeit definiert und gewichtet werden können und der Entwurf von 3. möglichst genauen, reproduzierbaren, manipulationssicheren und nachvollziehbaren Maßen sowie 4. von Algorithmen zur Berechnung nachvollziehbarer und reproduzierbarer Maße für technische Nachhaltigkeit auf Grundlage von Softwaremetadaten. Während 4. Gegenstand zukünftiger Forschungsvorhaben sein soll, können die Problemstellungen 1. bis 3. hier kurz detailliert werden.

Technische Nachhaltigkeit von Forschungssoftware

Während im Zusammenhang mit Software der Begriff der Nachhaltigkeit strukturell und inhaltlich ambig diskutiert wird (vgl. Tate 2005; Penzenstadler 2013; Goble 2014; Gröger & Köhn 2015), geben Becker et al. (2015) eine kurze Definition von technischer Nachhaltigkeit: "Technische [Nachhaltigkeit] bezieht sich auf die Langlebigkeit von Information, Systemen und Infrastruktur und deren angemessene Evolution unter sich ändernden Umgebungsbedingungen" (Übers. d. Autors). Konkretisiert man die Aspekte 'Langlebigkeit' und 'Evolution' auf Grundlage des Nachhaltigkeitsmodells von Jörissen et al. (1999), lassen sich drei Ziele technischer Nachhaltigkeit definieren: 1. Sicherung der Existenz der Software, 2. Erhaltung des Produktivpotentials der Software, 3. Schaffung und Bewahrung der Weiterentwicklungs- und Adaptionsmöglichkeiten der Software.

Kriterien

Das zu entwickelnde Metadatenrepositorium muss demnach Metadaten sammeln, die im Hinblick auf diese drei Ziele quantifizierbar sind. Voraussetzung dafür sind Kriterienkataloge, die sich jeweils auf eines der Ziele beziehen und auf Basis derer die Kategorisierung und Quantifizierung der Metadaten erfolgt. Für das Erstellen der Kriterienkataloge bieten etwa Jackson, Crouch & Baxter (2011), CodeMeta (codemeta.github.io), oder OntoSoft (ontosoft.org) eine Arbeitsgrundlage, die jedoch durch dort fehlende Kriterienkategorien wie beispielsweise 'personelle Mittel' erweitert werden muss, etwa auf dem Wege des Crowdsourcing.

Interaktivität

Weiterhin liegt es nahe, Teile des Repositoriums interaktiv zu gestalten. Dies dient nicht nur der Bindung der Community an das Werkzeug, sondern erlaubt vor allem die Gewinnung zusätzlicher Metadaten - beispielsweise zu tatsächlicher und dokumentierter Nutzung einer Forschungssoftware -, sowie die Evaluation bereits hinterlegter.

Maße für technische Nachhaltigkeit

Direkte Interaktion von Nutzern allerdings gefährdet massiv die Objektivität und Reproduzierbarkeit eines zu errechnenden Maßes für die technische Nachhaltigkeit einer Forschungssoftware. Daher ist eine Verteilung der Bemessung über mehrere Maße geraten: Einfach zu quantifizierende und objektive Metadaten (beispielweise Offenheit des Quellcodes, interaktiv hinterlegte Verwendungsnachweise) tragen zu einem harten Maß bei, objektive aber weniger einfach zu quantifizierende Metadaten (beispielsweise das verwendete Buildsystem) zu einem mittelharten Maß, weitere interaktiv erhobene, subjektive und qualitative Metadaten (beispielsweise Nutzerfreundlichkeit) zu einem

weichen Maß. Während die härteren Maße auf Grund ihrer Objektivität reproduzierbar und verhältnismäßig manipulationssicher sind, bietet das weiche Maß interpretationswürdige Anhaltspunkte.

Ausblick

Das beschriebene Metadatenrepositorium wäre ebenfalls ein geeignetes Instrument für die Dokumentation neu erstellter Forschungssoftware in Softwaremanagementplänen.

Eine vollständige Entwicklung des umrissenen Konzeptes und seine Implementierung ist im Rahmen der Dissertation des Autors geplant.

Bibliographie

Becker, Christoph / Chitchyan, Ruzanna / Duboc, Leticia / Easterbrook, Steve / Penzenstadler, Birgit / Seyff, Norbert / Venters, Colin C. (2015): „Sustainability design and software: The Karlskrona Manifesto“, in: *IEEE/ACM 37th IEEE International Conference on Software Engineering* 467–476.

Druskat, Stephan (2016): „Lightning Talk: A Proposal for the Measurement and Documentation of Research Software Sustainability in Interactive Metadata Repositories“, in: *Proceedings of the Fourth Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE4)* http://ceur-ws.org/Vol-1686/WSSSPE4_paper_20.pdf [letzter Zugriff 1. Dezember 2016].

Goble, Carole (2014): „Better software, better research“, in: *IEEE Internet Computing* 18: 4–8.

Gröger, Jens / Köhn, Marina (2015): „Nachhaltige Software. Dokumentation des Fachgesprächs Nachhaltige Software am 28.11.2014“, in: *Umweltbundesamt, Dokumentationen* 07/2015 http://www.umweltbundesamt.de/en/publikationen/nachhaltige-software [letzter Zugriff 31. August 2016].

Hettrick, Simon (2016): *Research software sustainability: Report on a Knowledge Exchange workshop* http://repository.jisc.ac.uk/6332/ [letzter Zugriff 31. August 2016].

Jackson, Mike / Crouch, Steve / Baxter, Rob (2011): *Software evaluation: criteria-based assessment*. Software Sustainability Institute.

Jörissen, Juliane / Kopfmüller, Jürgen / Brandl, Volker / Paetau, Michael (1999): *Ein integratives Konzept nachhaltiger Entwicklung*. Karlsruhe: Forschungszentrum Karlsruhe.

Penzenstadler, Birgit (2013): „Towards a definition of sustainability in and for software engineering“, in: *Proceedings of the 28th Annual ACM Symposium on Applied Computing* 1183–1185.

Tate, Kevin (2005): *Sustainable Software Development: An Agile Perspective*. Boston, MA: Addison-Wesley.

Living Books about History

Baumann, Jan
jan.baumann@infoclio.ch
infoclio.ch, Schweiz

Kurmann, Eliane
eliane.kurmann@infoclio.ch
infoclio.ch, Schweiz

Natale, Enrico
enrico.natale@infoclio.ch
infoclio.ch, Schweiz

infoclio.ch hat 2016 das digitale Projekt *Living Books about History* lanciert. Die Living Books sind eine neue Form digitaler Anthologien. Sie präsentieren kurze Essays zu aktuellen wissenschaftlichen Themen, die von ausgewählten online und frei verfügbaren Beiträgen begleitet werden. Das Projekt erprobt ein neues Format der wissenschaftlichen Publikation und will mit dem Wiederentdecken und Neuverwenden wissenschaftlicher Texte und Quellen auf die Chancen von Open Access aufmerksam machen.

Auf einem Poster soll das Konzept des Projekts kurz erläutert und die sechs bereits online verfügbaren Living Books vorgestellt werden. Dem Tagungsthema entsprechend werden dabei auch jene Aspekte des Projekts beschrieben, die die technische Nachhaltigkeit der Webseite und die konzeptionellen Ziele, die auf eine langfristige Nutzung digital verfügbarer Inhalte ausgerichtet sind, darlegen. Sie finden kurze Ausführungen zur Gestaltung des Posters im zweiten Teil dieser Bewerbung.

Die *Living Books about History* passen in mehrfacher Hinsicht ausgezeichnet zum Tagungsthema „Digitale Nachhaltigkeit“:

- In konzeptueller Hinsicht verfolgt das Projekt die Idee, besonders lesenswerte oder zu Unrecht vergessen gegangene Texte aus der Masse der im Internet verfügbaren Informationen hervorzuheben. Mit dem Hervorheben sollen relevante Ressourcen und herausragende wissenschaftlicher Beiträge auch im digitalen Raum langfristig sichtbar und einfach zugänglich bleiben.
- Das digitale Projekt ist auch in technischer Hinsicht auf Nachhaltigkeit bedacht: Zum einen wird durch die Vergabe von Digital Object Identifiers (DOI) für jedes Living Book sichergestellt, dass die verlinkten Webseiten auch bei einer Veränderung der URL erreichbar bleiben; zum andern wird durch das Archivieren der Website im Webarchiv Schweiz durch die Schweizerische Nationalbibliothek garantiert, dass die Living Books dauerhaft zugänglich sind.
- Inhaltlich geht insbesondere das von Tara Andrews herausgegebene Living Book „Digital Humanities“ der Geschichte dieses Fachs nach. Auch die anderen Living Books beschäftigen sich in formaler Hinsicht mit dem Digitalen, in dem u.a. vielseitige Quellenformate wie Videos, Webseiten oder Bilder in die jeweiligen Anthologien integriert werden.
- Mit der Sensibilisierung für die juristischen Bestimmungen wird die Nutzung online verfügbarer Beiträge gefördert. Das Projekt verweist bei allen Beiträgen auf die bibliografischen Referenzen der Erstveröffentlichung und die Nutzungsbedingungen.

Weitere Informationen zum Projekt sowie zum Design, an das wir uns bei der Ausgestaltung des Posters anlehnen würden, finden Sie unter: http://www.livingbooksabouthistory.ch/de/

In der Gestaltung des Posters sind folgende Punkte vorgesehen:

1.) Kurze Einführung ins Projekt:

- Konzept und Ziele
- Nachhaltigkeit der digitalen Infrastruktur
- Open Access und Nutzungsrechte

2.) Präsentation der online und frei zugänglichen Living Books:

- Tara Andrews - Digital Humanities: Diese Anthologie gibt eine Einführung in die Digital Humanities. Im Fokus stehen die Geschichte und die Begriffe sowie Ratschläge zum Einstieg in die Digital Humanities.
- Almut Höfert – Wunder und Monster im Mittelalter: Das Living Book beschäftigt sich mit Wundern und Mirakeln sowie der gesellschaftlichen Bedeutung, die ihnen im Mittelalter zukam.
- Guido Koller & Sebastian Schüpbach – Geschichte der modernen Verwaltung: Quellen und Berichte aus dem Schweizerischen Bundesarchiv geben Einblick in die Entwicklung der Verwaltung im 19. und 20. Jahrhundert.
- Martin Lengwiler & Beat Stüdli – Geschichte des Wohlfahrtstaats: Die Anthologie gibt einen Einblick in verschiedene Modelle des Wohlfahrtstaats und beschäftigt sich mit der Entwicklung, die zur heutigen Vielfalt geführt hat.
- Daniel Speich Chassé – „La situation coloniale“: In diesem Living Book geht es um Nord-Süd-Beziehungen im 20. Jahrhundert. Der Ausgangspunkt bildet ein Text von Georges Balandier aus dem Jahr 1951.

- Valérie Schafer – "Histoires de l'Internet et du Web" (ab Herbst 2016): Anhand einer Auswahl von Quellen und Aufsätzen wird die Geschichte des Internets und des Webs nachgezeichnet.

3.) Hinweis darauf, dass die Reihe fortgesetzt wird und wir Themenvorschläge für neue Living Books gerne entgegennehmen.

Maßnahmen zur digitalen Nachhaltigkeit in Langzeitprojekten – Das Beispiel *Capitularia*

Schulz, Daniela
schulzd1@uni-koeln.de
Bergische Universität Wuppertal; Universität zu Köln

Fischer, Franz
franz.fischer@uni-koeln.de
Cologne Center for eHumanities

Geißler, Nils
nils.geissler@uni-koeln.de
Cologne Center for eHumanities

Gödel, Martina
mgoedel@uni-koeln.de
Cologne Center for eHumanities

Bei der Planung und Durchführung von langfristigen (Editions-)Projekten stellen sich hinsichtlich der Nachhaltigkeit besondere Herausforderungen, da sich die technischen Entwicklungen der nächsten Jahre und Jahrzehnte eben nur bedingt vorausahnen lassen. Gerade in der Anfangsphase müssen aber bereits zahlreiche, oftmals richtungsweisende Entscheidungen getroffen werden, die den zukünftigen Erfolg oder Misserfolg, potenziell auftretende Probleme und Lösungsmöglichkeiten determinieren. Allerdings scheint es auch überhaupt nur in Projekten mit langer Laufzeit möglich, jenseits von reinen Willensbekundungen umfassende Strategien zur Nachhaltigkeit zu entwickeln und entsprechende Maßnahmen zu ergreifen. Damit leitet sich gleichzeitig die Pflicht ab, dies auch zu tun.

Am Beispiel des Projektes *Capitularia*, einem Langzeitprojekt (2014-2029), welches mit einer Hybrid-Edition frühmittelalterlicher Herrschererlasse befasst ist, sollen die verschiedenen Ebenen digitaler Nachhaltigkeit, damit verbundene Herausforderungen sowie erste Lösungsansätze präsentiert werden. Der von uns vorgeschlagene und in Teilen bereits umgesetzte Maßnahmenkatalog betrifft die Ebenen

- Datenmodellierung und Textauszeichnung,
- Datenhaltung und Dokumentation,
- Infrastrukturen (technisch und institutionell),
- Webseite und verwendete (Web-)Technologien,
- Präsentation, Zugänglichkeit und Nachnutzbarkeit der Forschungsergebnisse.

Bei der **Textauszeichnung** sollte auf etablierte Standards zurückgegriffen werden, um die programm- und plattformunabhängige Weiterverarbeitung der Daten langfristig zu gewährleisten. Bei *Capitularia* werden Transkriptionen, Handschriftenbeschreibungen, Register und bibliographische Daten gemäß der im Projekt erarbeiteten Richtlinien in *TEI-XML* codiert und durch ein den projektspezifischen Anforderungen angepasstes, restriktives Schema sowie den Einsatz von *Schematron* (ISO/IEC 19757-3) kontrolliert, um damit über die gesamte Projektlaufzeit hinweg und auch bei wechselndem Personal die Konsistenz und Einheitlichkeit und damit die Qualität der erstellten Daten zu sichern. (Hedler u.a. 2011: 11-12)

Dies bedingt von Beginn an eine **umfassende Dokumentation**, die nicht nur z.B. in Form eines Wikis vor allem für interne Zwecke verwendet wird, sondern auch möglichst viele Informationen öffentlich zugänglich macht, so dass andere Projekte von den Erfahrungen profitieren können. Die Entwicklung und Rationalisierung von Arbeitsprozessen sowie deren genaue Darlegung gewährleisten dabei neuen Mitarbeitern einen möglichst einfachen Einstieg. Für die öffentliche Dokumentation bieten sich neben der eigenen Projektwebseite beispielsweise Dienste wie *GitHub* an, auf denen gleichzeitig eine **transparente Datenhaltung** mit Versionierung möglich ist, und auch eigene technische Entwicklungen einfach zur Nachnutzung bereitgestellt werden können. Dies alles setzt natürlich die Verwendung entsprechender *Open Access* Lizenzen voraus.

Die **Sicherung der Langzeitverfügbarkeit** einer Ressource lässt sich generell nur durch eine entsprechende **technische Infrastruktur** (Daten-, Kompetenz-, Rechenzentren) in Kombination mit **institutioneller Anbindung** (Universitäten, Forschungsbibliotheken, kulturbewahrende Institutionen) gewährleisten. (Wissenschaftsrat 2011) Zu unterscheiden ist dabei zwischen der langfristigen Archivierung und Vorhaltung der Forschungsdaten und dem möglichst langen Erhalt der Webressource insgesamt, die ja ebenfalls mit allen Funktionalitäten über die Projektlaufzeit hinaus verfügbar sein soll. (Oßwald u.a. 2012: 13-15) Die Voraussetzungen in Köln erscheinen ideal, da dem Projekt mit dem *CCeH* ein im Bereich der DH ausgewiesener technischer Partner zur Verfügung steht, der durch enge Kooperation mit weiteren universitären Einrichtungen wie dem *Data Center for the Humanities* und dem Kölner Rechenzentrum, ideale strukturelle Voraussetzungen für Aufbau und langfristige Erhaltung digitaler Ressourcen schaffen konnte. Zusätzlich nimmt das *Capitularia*-Projekt auch am Webarchivierungsprogramm der Bayerischen Staatsbibliothek teil.

Für die Webpräsentation wurden in enger Zusammenarbeit mit dem technischen Partner **geeignete (open source) Technologien** ausgewählt, deren zukünftiger Erhalt sowie deren Weiterentwicklung von einer breiten Community getragen werden, und die sich in anderen Projekten bereits bewährt haben. Im Fall von *Capitularia* wird aktuell das PHP-basierte Content Management System *WordPress* zur Verwaltung der Webpräsenz verwendet. Dabei werden die Standardfeatures (Blog, Suche, Mehrsprachigkeit etc.) möglichst breit genutzt und vereinzelt um weitere Eigenentwicklungen ergänzt, die insbesondere den Bereich des XSLT-Pipelining betreffen. Diese im Kontext des Projektes entwickelten *Plugins* sollen auch der Nachnutzung durch andere Projekte zur Verfügung stehen. WordPress erfüllt durch die freie Zugänglichkeit des Quellcodes, der aktiven und heterogenen Entwickler-Gemeinschaft, des Ökosystems, welches sich um diese Software etabliert hat, sowie der verwendeten Lizenzmodells wichtige Voraussetzungen digitaler Nachhaltigkeit. (Stürmer 2015, S. 36-37) Die bewusst gewählten *low-tech*-Lösungen gewährleisten weiterhin langfristig die leichte Wartbarkeit des Systems sowie plattform- und auch personelle Unabhängigkeit. Bei Bedarf könnte somit ohne größere Probleme auf ein anderes Präsentationsframework umgezogen werden.

Die **Präsentation der Forschungsergebnisse** findet auf mehreren Ebenen statt. Die angefertigten Transkriptionen werden zusammen mit weiteren Materialien projektbegleitend auf der Webseite veröffentlicht. Die Inhalte sind dort über Permalinks adressierbar und die zugrunde liegenden Daten (*XML*) stehen zum Download bereit. Um die Nachnutzung der Daten weiter zu erleichtern, ist die Implementierung von Schnittstellen (z.B. REST) vorgesehen. [1] Die kritische Edition erscheint in Druckfassung in der *Leges*-Reihe der *Monumenta Germaniae Historica* und wird somit auch langfristig über deren Online-Angebot (dMGH) verfügbar sein. Vorabversionen der kritischen Editionstexte werden aber bereits zeitnah in digital aufbereiteter Form auf der Webpräsenz zur Verfügung gestellt.

Wenn auch finanzielle Ausstattung und Laufzeit von Projekten und damit auch deren Handlungsmöglichkeiten hinsichtlich Nachhaltigkeitsstrategien durchaus unterschiedlich sind, so erscheinen doch die Bereiche, in denen Maßnahmen getroffen werden können, allgemeingültig zu sein. Mit den hier vorgestellten Ansätzen soll ein aktiver und vor allem praxisorientierter Beitrag zur Diskussion um digitale Nachhaltigkeit geleistet werden.

Fußnoten

1. Die freie Bereitstellung von im Rahmen von Digitalen Editionen entstandenen (XML-)Daten lässt aktuell noch zu wünschen übrig. Nur durch diese wird aber Nachnutzung und damit integrative und innovative Forschung erst ermöglicht. (Turska u.a. 2016: 1) Technische Schnittstellen können die Nachnutzung befördern und stellen daher ein Kriterium einer aktuellen Ansprüchen genügenden digitalen Edition dar. (Sahle u.a. 2014).

Bibliographie

Hedler, Marko / Montero Pineda, Manuel / Kutscherauer, Nico (2011): *Schematron. Effiziente Business Rules für XML-Dokumente*. Heidelberg: dpunkt-Verlag.

Oßwald, Achim / Scheffel, Regine / Neuroth, Heike (2012): „Langzeitarchivierung von Forschungsdaten. Einführende Überlegungen“, in: Neuroth, Heike / Strathmann, Stefan / Oßwald, Achim / Scheffel,

Regine / Klump, Jens / Ludwig, Jens (eds.): *Langzeitarchivierung von Forschungsdaten: Eine Bestandsaufnahme*. Boizenburg: Verlag Werner Hülsbusch 13–23 http://www.nestor.sub.uni-goettingen.de/bestandsaufnahme/nestor_lza_forschungsdaten_bestandsaufnahme.pdf [letzter Zugriff 9. November 2016].

Sahle, Patrick / Vogeler, Georg / IDE (eds.) (2014): *Kriterien für die Besprechung digitaler Editionen*. Version 1.1 Juni 2014 http://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/ [letzter Zugriff 12. November 2016].

Stürmer, Matthias (2015): „Wann sind Open Source Projekte digital nachhaltig?“, in: swissICT und Swiss Open Systems User Group (eds.): *Open Source Studie: Schweiz 2015* 36–37 http://www.swissict.ch/fileadmin/customer/Publikationen/OSS-Studie2015.pdf [letzter Zugriff 9. November 2016].

Turska, Magdalena / Cummings, James / Rahtz, Sebastian (2016): „Challenging the Myth of Presentation in Digital Editions“, in: *Journal of the Text Encoding Initiative* 9 http://jtei.revues.org/1453 [letzter Zugriff 12. November 2016].

Wissenschaftsrat (ed.) (2012); *Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsstrukturen in Deutschland bis 2020*. Drs. 2359-12. Berlin http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf [letzter Zugriff 09. November 2016].

maus - eine WebApp zur einfachen Erstellung funktionaler Webdokumente

Dufner, Matthias
matthias.dufner@hs-mainz.de
Hochschule Mainz, Deutschland

Kunz, Axel
axel.kunz@hs-mainz.de
Hochschule Mainz, Deutschland

Klammt, Anne
klammt@mainzed.org
Mainzer Zentrum für Digitalität in den Geistes- und Kulturwissenschaften (mainzed), Deutschland

Oft kommen Fachwissenschaftler(innen) in die Lage, Texte für zahlreiche verschiedene digitale Formate (Blogs, Präsentationen, E-LeraningPlattformen) zu erzeugen. Sie stehen dabei vor der Wahl, Texte immer wieder neu einzupassen oder sie in Dateiformaten zur Verfügung zu stellen, die strenggenommen zweckentfremdet werden. Powerpoint-Präsentationen sind als begleitendes Element zu einer Vorlesung gedacht und nicht als Lernunterlagen, eine PDF-Datei soll im ursprünglichen Sinne eine Druckvorstufe sein, dementsprechend sind sowohl Usability als auch Lesbarkeit auf kleineren Bildschirmen meistens mangelhaft. HTML5 Dateien kennen diese Probleme nicht. Sie sind gewissermaßen digitale Rohdaten, die ohne zusätzliche Software von jedem modernen Browser interpretiert und dargestellt werden können. In HTML zu schreiben ist jedoch unkomfortabel und erfordert Kenntnisse. maus setzt genau an dieser Stelle an und bietet eine leicht nutzbare Entwicklungsumgebung, die reichhaltige HTML5-Dokumente exportieren kann. Dank der Trennung von Inhalt und Formatierung können dabei alle mit maus erzeugten Dokumente in verschiedenen Kontexten wiederverwendet werden. Daher wird maus vom Mainzer Zentrum für Digitalität in den Geistes- und Kulturwissenschaften (mainzed) aktuell auch besonders mit Blick auf die Erzeugung nachhaltiger, offener digitaler Lehrmaterialien (OER) weiterentwickelt.

maus ist eine neu entwickelte Webanwendung. Sie besteht aus einem Editor, indem mit einfachem Markdown Texte strukturiert und semantisch angereichert werden. Diese Texte werden in einem weiteren Schritt in HTML5 überführt und können automatisiert mit CSS-Templates verknüpft werden. Die vereinfachte Auszeichnungssprache Markdown fungiert somit als technische Brücke. Die Markdown-Syntax ist leicht zu erlernen und belässt Inhalte in lesbarer Form – somit bleiben die Daten trotz Auszeichnungen übersichtlich und klar. Tatsächlich nutzt Markdown einfache Auszeichnungselemente, um komplexe Strukturen im HTML zu generieren. Die erstellten Dokumente können später als HTML-Dateien exportiert werden. Der Editor unterstützt Syntax-Highlightning und eine Vorschau des Dokuments. Er ist ansonsten bewusst einfach und schlank gehalten. maus unterscheidet sich damit auch von der Usability erheblich von den üblichen CMS-Systemen. Dennoch bietet auch maus eine Verwaltung von Nutzer(inne)n und Nutzergruppen. Weiterhin besteht die Möglichkeit, private Dokumente zu

erstellen oder Inhalte mit anderen Nutzern zu teilen. Ebenso ist das Zurücksetzen auf frühere Versionen eines Dokuments möglich.

Die Entwicklung ist aber noch einen Schritt weitergegangen, denn oftmals reichen die durch Markdown unterstützen Auszeichnungselemente nicht aus. Hierzu zählen beispielsweise das automatische Generieren von Inhalts- und Quellenverzeichnissen, komplexen Bildunterschriften oder dynamische Begriffserläuterungen. Um die hierfür notwendigen komplexen HTML-Strukturen automatisiert zu generieren, wurden neue Auszeichnungselemente eingeführt. Hierdurch können die Markdown Dokumente u. a. mit semantischen Elementen angereichert werden. Beispielsweise wird durch die Auszeichnung: {definition: mainzed} ein HTML-Element erzeugt, das später im HTML-Dokument bei eine Erklärung des Begriffes 'mainzed' einblendet sobald sich der Cursor über dem Element befindet. Genauso kann maus aber auch Markdown-Dokumente ausgeben und hierbei die Erweiterungen entfallen lassen. Die Erweiterungen des Markdown sind also Optionen, die die Dokumente nicht korrumpieren.

Die Stärke von maus liegt in der Möglichkeit, erstellte Bausteine je nach Anwendungszweck zu passgenauen Medien zusammensetzen zu können. Ihre Darstellung lässt sich mit Hilfe unterschiedlicher Templates (siehe Templates) ebenfalls ändern. Mit Hilfe von maus lassen sich somit aus den selben digitalen Rohdaten problemlos Lesedokumente, Präsentationen oder Websites entwickeln. Die Arbeitsweise ist dabei grundsätzlich auf Nachhaltigkeit angelegt, weil Inhalt und Gestaltung getrennt bleiben. Es ist sehr einfach möglich, einmal erstellte Inhalte zu einem späteren Zeitpunkt in anderer Form wiederzuverwenden.

Bei der Software Architektur handelt es sich um einen MEAN Stack, einem Paket an freier Open-Source-Software zur Entwicklung von dynamischen Webanwendungen. Es besteht aus MongoDB, Express, AngularJS und Node.js. Weiterhin wird CodeMirror genutzt, ein JavaScript basierender Texteditor, der die Hervorhebung von Markdown-Auszeichnungselementen unterstützt. Dieser wurde erweitert, um auch die neu erzeugten Elemente hervorzuheben. Der Markdown parser und compiler marked wird zum Konvertieren der Markdown- in HTML-Dokumente verwendet. maus selbst ist, wie seine einzelnen Komponenten als Open-Source-Projekt auf GitHub (https://github.com/mainzed/maus) freigegeben und kann somit frei genutzt und weiterentwickelt werden.

maus wird bereits in Projekten des mainzed eingesetzt, um beispielsweise Lehrmaterialien für die Lernplattform OpenOLAT oder Inhalte für den mainzed-Jahresbericht durch Fachwissenschaftler(innen) erstellen zu lassen.

Im Fokus der Weiterentwicklung stehen derzeit die vereinfachte Erweiterbarkeit der Anwendung durch erfahrene Nutzer(innen). Diese sollen in der Lage sein, eigene Layouts und Anreicherungselemente anzulegen, um die momentan verfügbaren Vorlagen zu erweitern. Denkbar wären Layouts für wissenschaftliche Poster, Paper, Präsentationen und auch Printmedien.

Nachhaltigkeit durch Zusammenschluss: Die DARIAH Data Re-Use Charter

Baillot, Anne
anne.baillot@gmail.com
Centre Marc Bloch, Deutschland

Busch, Anna
anna_busch@gmx.de
Forschungsverbund Weimar Marbach Wolfenbüttel

Puren, Marie
marie.puren@inria.fr
Inria Paris

Mertens, Mike
mike.mertens@dariah.eu
Goettingen Center for Digital Humanities

Romary, Laurent
laurent.romary@inria.fr
Inria Paris

Bedarf und Forschungsstand

Ausgangspunkt der Charter ist die Feststellung, dass es bis dato keinen klaren Rahmen gibt, der die wissenschaftlichen Weiternutzungsbedingungen von digitalen

Daten regelt, die auf Beständen von Kulturerbeinstitutionen basieren. Einige allgemeine Richtlinien wurden entwickelt, die allerdings nicht spezifisch die Interaktion zwischen digitalen Kulturerbedaten und Forschung in den Blick nehmen, sondern breiter angelegt sind. So etwa die Data Re-Use Models DANS oder die UNESCO Guidelines for the preservation of digital heritage (die wohlgemerkt nicht ausschliesslich Kulturerbedaten in den Blick nehmen). Andere Initiativen haben die Evaluation der digitalen Best-Practices in den Mittelpunkt gestellt, so etwa der Data Seal of Approval oder die Richtlinien zur Certification and Assessment of Digital Repositories des Center for Research Libraries. Im Bereich des Forschungsökosystems selbst jedoch kann in diesem Sinne hauptsächlich auf den CERN Code of Conduct hingewiesen werden, der zwar wenig konkrete Verpflichtungen definiert, dafür aber den Schwerpunkt auf eine Ethik der wissenschaftlichen Zusammenarbeit legt, dessen Geist ebenfalls in der hier dargestellten Charter von wesentlicher Bedeutung ist.

Der Mangel eines klaren Rahmens für die Zusammenarbeit an und mit digitalen Kulturerbedaten macht sich für alle beteiligten Akteure im Alltag bemerkbar: Forscher und Forscherinnen, die vor einem Scan stehen, ohne zu wissen, wie sie diesen weiterverwenden und zitieren dürfen; Kulturerbeeinrichtungen, die ihre Metadatensätze mit den entstehenden Forschungsarbeiten verknüpfen möchten; Equipments, die Anfragen für die gleichen Artefakte immer wieder bekommen; Datenzentren, die der Schnittstelle zwischen Kulturerbeeinrichtungen und Forschung fernbleiben bzw. von der Verwaltung der unterschiedlichen Rechte der von ihnen gehosteten Daten herausgefordert sind; Forschungseinrichtungen, denen Strukturen fehlen, um ihren Mitarbeitern und Mitarbeiterinnen good-practice-Empfehlungen an die Hand zu geben. Der Mangel an definitorischer Schärfe des betroffenen Informationsaustauschs steht im Gegensatz sowohl zu dem damit für alle Beteiligten zusammenhängenden Bedarf als auch zu dem offensichtlichen Vorteil, den ein solcher für das ganze Ökosystem repräsentiert. Nachhaltig kann ein solcher Austausch nur dann werden, wenn neben den Datenproduzenten (Forscher/innen, Kulturerbeeinrichtungen, Equipments) auch Datenzentren daran beteiligt sind, wie es hier vorgesehen ist. Einige Einrichtungen übernehmen zwar mehrere dieser Funktionen, aber die Kommunikation zwischen den entsprechenden Abteilungen ist auch dort nicht immer optimal.

Nicht zuletzt für die Weiterentwicklung der Digital Humanities handelt es sich bei dem hier geschilderten Zusammenhang um ein zentrales Anliegen, bilden diese Daten ja die Grundlage für eine beachtliche Reihe geisteswissenschaflicher Forschungen etwa in den Bereichen der Archäologie, der Kunstgeschichte, der Musikwissenschaft, aber auch u.a. der Literaturgeschichte und der Editionswissenschaft. Selbst vom Standpunkt nicht-historischer Fächer ist es auf Dauer von Vorteil, wenn für digitale Daten aller Art hinsichtlich der Formate und Standards Rücksprache gehalten wird, wenn Equipments die bereits durchgeführten Scanarbeiten transparent machen, wenn auf einen Blick klar werden kann, zu welchen Kulturerbebeständen in welchen Datenzentren Arbeiten vorhanden sind. Die DARIAH Data Re-Use Charter hat zum Ziel, die Transaktionen zwischen all den Akteuren, die an der wissenschaftlichen Arbeit mit den digitalen Daten, die auf Kulturerbedaten basieren, Interesse haben, einfacher, transparenter und nachhaltiger zu machen, sodass alle daraus einen Nutzen für ihre eigene Arbeit ziehen können.

Grundprinzipien

Die DARIAH Data Re-Use Charter ist eine Interaktionsplattform für alle an der wissenschaftlichen Nutzung digitaler Daten von Kulturerbeeinrichtungen interessierten Akteure. Die Charter stellt diesen Textbausteine zur Verfügung, damit sie die Bedingungen ihrer Zusammenarbeit gemeinsam definieren können. Auf diesem Weg kann eine Kulturerbeinstitution die Nutzungsbedingungen ihres Gesamtbestands, aber auch von Sondersammlungen unterschiedlich beschreiben. Eine Universität kann ihre Befürwortung der Open Access-Prinzipien deklarieren und ihren Mitarbeiter/innen empfehlen, ihre Kooperationen mit Kulturerbeeinrichtungen in diesem Sinne zu konzipieren. Wie in diesen zwei Beispielen liefert die Charter ausformulierte Bausteine für die Zusammensetzung einer Kooperationsvereinbarung, an der ebenfalls Infrastruktureinrichtungen beteiligt sind, die mit der nachhaltigen Sicherung der Primär- bzw. Sekundärdaten betraut sind. Auf diese Art und Weise werden Empfehlungen in folgenden Bereichen formuliert (wobei mehrere

Formulierungen pro Bereich zur Verfügung stehen sowie jeweils ein freies Textfeld):

Zugang zu Metadaten, Texten, Bildern (beispielsweise im Fall einer zu edierenden Handschrift: Archivmetadaten, Transkription und Annotation, Scan, die dann auch explizit miteinander verlinkt werden)

- Lizenzierung der Inhalte (mit Verweis auf weiterführende, informierende Dokumentation)
- Formate und Standards (ebenso)
- Anreicherungen; Verknüpfung der wissenschaftlichen Anreicherungen der Kulturerbedaten mit den Metadaten
- Streuung sowohl der Kulturerbedaten als auch der Anreicherungen
- Qualitätssicherung bei allen beteiligten Akteuren

Darüber hinaus nennt jede Einrichtung den/die einschlägige(n) Ansprechspartner/in, damit eine Kommunikation erleichtert wird.

Neben den technischen Aspekten, bei denen mit Sicherheit Aufklärungsbedarf besteht, geht es auch, wenn nicht vorrangig, darum, eine digitale Kooperationsethik zu fördern, die wie im CERN Code of Conduct auf dem Respekt vor dem Werk anderer, der guten Zusammenarbeit, der Förderung von Kooperation und der Offenheit gegenüber der Öffentlichkeit beruht.

Umsetzung

Die Charter wird die Form eines Webinterfaces annehmen, auf dem sich die Akteure in ihrer jeweiligen Funktion (Forscher/in, Kulturerbeinstitution, Datenzentrum, Equipment, Forschungseinrichtung) registrieren lassen können und von dort ausgehend unter den ihnen zur Verfügung stehenden Optionen diejenigen aussuchen können, zu denen sie sich bekennen möchten. Im Vortrag wird nach einer Einführung zum Grundgedanken der Charter spezieller auf den Teil des Interfaces eingegangen, das dem Forscher/der Forscherin gewidmet ist.

Als registrierter Forscher/registrierte Forscherin soll man sich in drei Bereichen positionieren, die jeweils entweder im öffentlichen Profil oder im privaten Bereich (nur für die eigens ausgewählten Kooperationspartner - Kulturerbeinstitution, Datenzentrum, Equipment - zugänglich) erscheinen.

Der erste Bereich betrifft den Zugang zu den Daten. Dort verpflichtet sich der Forscher dazu, den Anforderungen der Kulturerbeinstitutionen zu folgen, was die Verwendung (insbes. Zitierweise) der digitalen Kulturerbedaten angeht. Hier kann der Forscher ebenfalls die Bestände anklicken, für die er sich interessiert und diese Information auch publik machen oder nur den betroffenen Institutionen zugänglich machen.

Der zweite Bereich betrifft die Streuung der vom Forscher auf der Grundlage der Kultuererbedaten produzierten Sekundärdaten oder angereicherten Metadaten. An dieser Stelle kann der Forscher die Lizenzen nennen, die er bevorzugt (weiterführende Dokumentation zu diesem Thema wird anklickbar gemacht). Eine Identifikation mittels einer ORCID-Nummer , die Referenzierung einer Id-HAL oder Vergleichbares können ebenfalls an dieser Stelle angegeben werden.

Im dritten Bereich geht es um den Umgang mit den anderen Charter-Partnern: Best-Practices wie die systematische Nennung der Kooperationspartner oder die explizite Nennung der Art und Weise, wie man selbst zitiert werden möchte, werden dort deklariert.

Diese drei Bereiche, die die Nutzung von Primärdaten, die Produktion- und Streuung von Sekundärdaten und die allgemeine Kooperationsethik bedingen und definieren, machen die Grundlage des Forscherprofils aus. Von dort ausgehend kann er dann die Einrichtungen, Bestände und Dienstleistungen recherchieren, mit denen er zusammenarbeiten möchte.

Die Charter hat somit eine doppelte Dimension: die einer Information- und Selbstpositionierungsplattform und die eines sozialen Netzwerkes, wobei auf die Balance zwischen Transparenz und Respekt der Privatsphäre stets geachtet wird ;

Perspektiven

Die DARIAH Data Re-Use Charter wird zwischen Herbst 2016 und Frühjahr 2017 soweit entwickelt sein, dass ein Soft Launch im Frühjahr 2017 stattfinden kann. Das Kernteam arbeitet sowohl an der Entwicklung der Webseite als auch an der Einholung von Feedback interessierter Akteure, um dieses in der Entwicklung des Interfaces im Allgemeinen und der relevanten Bausteine im Besonderen zu integrieren. Sie wurde bereits in ihren Ansätzen auf Konferenzen dargestellt und wird im Herbst in ähnlichen

Kontexten vorgestellt, damit weiteres Feedback eingeholt werden kann. Darüber hinaus findet im November 2016 in Berlin eine dedizierte Arbeitssitzung statt sowie eine Woche später eine entsprechende in Paris und im Januar eine in Rom - womöglich können über den Winter ebenfalls in anderen EU-Ländern entsprechende Sitzungen stattfinden. Bis zur DHd wird das Interface als das Ergebnis dieser breitgefächerten Konsultation demonstrierbar sein und kurz vor dem Launch stehen. Das bereits signalisierte Interesse zahlreicher Einrichtungen und Forscher/innen lässt vermuten, dass die Vernetzungsfunktion der Plattform rasch Konturen gewinnen wird.

Ausgerechnet dieser Aspekt gilt es mit Blick auf Nachhaltigkeit zu unterstreichen: Wenn Equipments systematischer zitiert werden, wenn Datenzentren expliziter an der Schnittstelle zwischen Kulturerbeinstitutionen und Forschung arbeiten können, wenn Forscher/innen ohne akademische Anbindung - und es gibt immer mehr prekäre Wissenschaftler/innen, die nichtsdestotrotz forschend tätig sind - einen Kooperationsraum finden können, dann ist für das gesamte Forschungsökosystem viel gewonnen und ein entscheidender Beitrag zur Nachhaltigkeit von digitalen Primär- und Sekundärkulturerbedaten geleistet.

Nachhaltigkeitperspektiven von Graphdaten

Kuczera, Andreas
andreas.kuczera@geschichte.uni-giessen.de
Regesta Imperii, Universität Gießen, Akademie der Wissenschaften Mainz, Deutschland

Beispiele Graphbasierter Erschließung der Datenbank Nomen et Gens

Die Datenbank Nomen et Gens ist aus einem DFG-Projekt hervorgegangen und verzeichnet Quellen und die in Ihnen belegten Personen für die vier Jahrhunderte vor der Zeit Karls des Großen. Das Abfragefrontend der Mysql-Datenbank ist im Internet unter www.nomen-et-gens.de zu erreichen.

Ein Teil der Datenbank wurde im Vorfeld eines DFG-Antrages in eine Graphdatenbank konvertiert um neue Auswertungsmöglichkeiten zu testen, die im Rahmen dieses Posters dargestellt werden sollen.

Frage an die Datenbank:

Zeige mir die Quelle „Annales Petaviani“ und alle Personen, die in dieser Quelle belegt sind und alle Quellen, in denen diese Personen wiederum gemeinsam belegt sind.

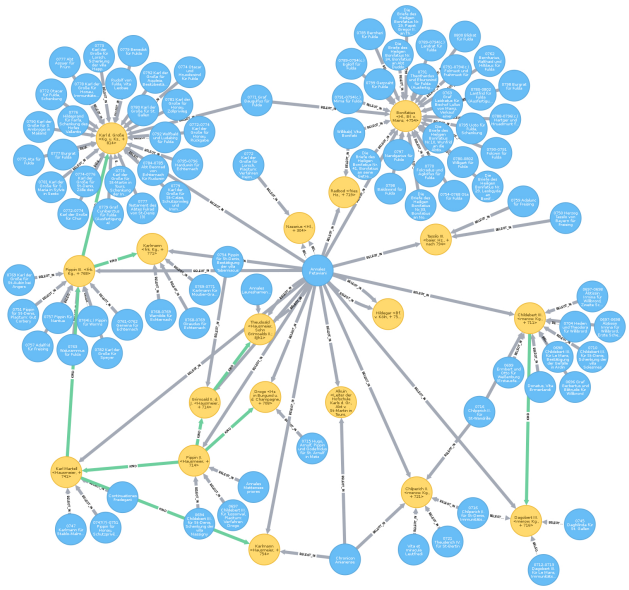


Abbildung 1: Visualisierung der Graphdatenbankabfrage.

Bewertung des Abfrageergebnisses:

Der interessanteste Information der Visualisierung ist die traversale Abfrage zu Personen in einer Quelle, die wiederum in einer Quelle gemeinsam vorkommen. Die Abfrage „Zeige mir alle Personen in einer Quelle" funktioniert ja auch mit einer relationalen Datenbank. Aber es gibt keine Möglichkeit, auf die Weise auch noch gleich zu sehen, in welcher Quelle eine Person außerdem noch steht. Insofern fügt die Graphdatenbank hier eine "Dimension" hinzu.

Verwandtschaftliche Beziehungen zwischen Personen stehen beim jeweiligen Personeneintrag und sind damit immer nur bis zum nächsten Glied zu sehen.

Für Frühmittelalterhistoriker besonders interessant sind gemeinsam urkundlich belegte Personen. Eine solche Abfrage ist über die relationale Datenbank nur schwer umfassend durchzuführen.

Die Visualisierung aus der Graphdatenbank veranschaulicht die Überlieferungssituation von einzelnen Personen. Der umkringelte Radbod ist, anders als z.B. Bonifatius zu

dem es sehr viele Belege gibt, nur schwer historisch fassbar.

Bibliographie

Kuczera, Andreas (2016c): „Digital Editions beyond XML – Graph-based Digital Editions“, in: *Proceedings of the 3rd HistoInformatics Workshop on Computational History (HistoInformatics 2016)* 37–46 http://ceur-ws.org/Vol-1632/paper_5.pdf .

Kuczera, Andreas (2016b): „Graphbasierte digitale Editionen“, in: *Mittelalter: Interdisziplinäre Forschung und Rezeptionsgeschichte* 19. April 2016 mittelalter.hypotheses.org/7994

Kuczera, Andreas (2016a): „Endcoding and Presenting Historical Biographical Data with Graph Data Bases“, in: *CO:OP. The Creative Archives' and Users' Network* https://coop.hypotheses.org/297 .

Kuczera, Andreas (2015): „Graphdatenbanken für Historiker. Netzwerke in den Registern der Regesten Kaiser Friedrichs III. mit neo4j und Gephi“, in: *Mittelalter: Interdisziplinäre Forschung und Rezeptionsgeschichte* 5. Mai 2015 mittelalter.hypotheses.org/5995

Kuczera, Andreas (2014b): „Big Data History“, in: *Mittelalter: Interdisziplinäre Forschung und Rezeptionsgeschichte* 10. Oktober 2014 mittelalter.hypotheses.org/3962

Kuczera, Andreas (2014a): „Digitale Perspektiven mediävistischer Quellenrecherche“, in: *Mittelalter: Interdisziplinäre Forschung und Rezeptionsgeschichte* 18. April 2014 mittelalter.hypotheses.org/3492

PaLaFra – Entwicklung einer Annotationsumgebung für ein diachrones Korpus spätlateinischer und altfranzösischer Texte

Döhling, Lars
lars.doehling@ur.de
Universität Regensburg, Deutschland

Burghardt, Manuel
manuel.burghardt@ur.de
Universität Regensburg, Deutschland

Wolff, Christian
christian.wolff@ur.de
Universität Regensburg, Deutschland

Ziel von *PaLaFra* [1] („Le passage du latin au français“) ist der Aufbau eines digitalen Korpus spätlateinischer und altfranzösischer Texte, das durch die Kombination von Lemmatisierung, syntaktischer und morphologischer Annotation sowie diskurspragmatischen und texttypologischen Deskriptoren komplexe Abfragestrategien ermöglicht und so eine qualitativ neuartige Nutzung der Texte bei der Rekonstruktion des lateinisch-romanischen Sprachwandels erreichen soll. Daran arbeitet ein deutsch-französisches Team der Universität Regensburg, der Universität Tübingen, der *École Normale Supérieure* in Lyon und der Universität Lille, das seit Sommer 2015 von der Deutschen Forschungsgemeinschaft (DFG) und der *Agence Nationale de Recherche* (ANR) gefördert wird. Das Projektteam ist interdisziplinär ausgerichtet und besteht aus romanischen Sprachwissenschaftlern, Computerlinguisten und Medieninformatikern. Während für den Bereich des Altfranzösischen auf das bestehende *Base de Français Médiéval*-Korpus [2] zurückgegriffen werden kann, so ist die Erstellung eines — was die Annotation angeht — kompatiblen Korpus spätlateinischer Texte ein wichtiges Teilziel des *PaLaFra*-Projekts.

In diesem Posterbeitrag berichten wir über Herausforderungen und Lösungsansätze bei der Erstellung einer Annotationsumgebung und eines diachronen Tagsets, das gleichermaßen in der Lage ist, die Idiosynkrasien der beiden Sprachstufen adäquat abzubilden, aber auch die diachronen Elemente im Sprachwandel einheitlich zu markieren.

Bereits für das spätlateinische Teilkorpus zeigt sich, dass es an einem standardisierten Tagset fehlt. Mindestens drei Varianten wurden in der Vergangenheit für die Annotation (spät-)lateinischer Texte entwickelt: *CoLaMer* (Selig et al. 2015), *CompHistSem* (Eger et al. 2015) und *LASLA* (Denooz 1978). Diese unterscheiden sich sowohl in den zugrunde liegenden linguistischen Konzepten als auch in ihrer Granularität. Demzufolge existiert auch kein einfaches Mapping zwischen ihnen. Für die Entwicklung eines sprachübergreifenden Tagsets in *PaLaFra* kommt erschwerend hinzu, dass die beiden Zielsprachen — Spätlatein und Altfranzösisch — trotz ihrer Verwandtschaft klare strukturellen Unterschiede aufweisen.

Zumindest für die Ebene der Wortarten (PoS, Part-of-Speech) liefert beispielsweise das Projekt *Universal Dependencies* [3] wichtige Anhaltspunkte für ein sprachübergreifendes Tagset. Dieses Projekt hat sich die Entwicklung sprachübergreifend-kompatibler Baumbanken als Ziel gesetzt hat, die auf universellen Wortartkategorien basieren. Trotzdem bedingt die Entwicklung eines übergreifenden Tagsets oft den manuellen Vergleich von Annotationen, z.B. durch visuelle Gegenüberstellung annotierter Parallelkorpora. Unsere Recherche ergab, dass es an einem adäquaten Werkzeug für diese Aufgabe mangelt. Einerseits gibt es unzählige Annotationswerkzeuge, welche auf die Darstellung nur eines Textes samt Annotationen fokussieren (Burghardt 2014, Neves and Leser 2014). Auf der anderen Seite gibt es Alignierwerkzeuge, die auf die parallele Darstellung von Texten spezialisiert sind, aber dabei Annotation meist ignorieren, z.B. *LF Aligner* [4], *Moses* [5] oder *ParaVoz* [6]. Um diese Lücke zu schließen, haben wir auf der Basis von *InterText* [7] — einem im Webbrowser zu bedienenden Alignierwerkzeug (Vondricka 2014) — ein Vergleichswerkzeug für annotierte Parallelkorpora entwickelt. Unsere Erweiterung unterstützt sowohl die Hervorhebung zueinander kompatibler (PoS-)Tags als auch die flexible Darstellung von Lemmata und morpho-syntaktischen Annotationen (Abbildung 1). Die dafür nötigen Informationen werden beim Import aus den TEI-XML-Daten extrahiert und mit Hilfe von JavaScript dynamisch visualisiert.

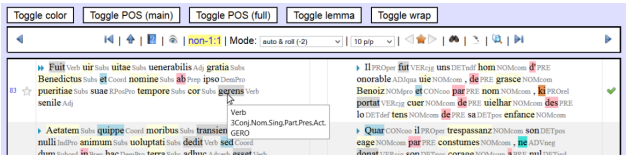


Abbildung : Das Bildschirmfoto zeigt die modifizierte InterText-Ansicht, erkennbar oben an der zusätzlichen Schalterleiste. Links ist die lateinische „Vita Benedicti“ (Vogüe und Antin 1979) zu sehen, annotiert mit dem LASLA Tagset (Denooz 1978), rechts das französische Gegenstück „Vie de saint Benoit“ (Foerster 1876), annotiert mit dem Cattex Tagset (Guillot et al. 2010). Aktuell ist sowohl die Hervorhebung kompatibler PoS-Tags („Toogle color“) als auch die Anzeige der vollständigen PoS-Annotation („Toogle POS (full)“) aktiviert.

Neben der eigentlichen Datenaufbereitung ist auch die Optimierung des Annotationsworkflows mit geeigneten Werkzeugen im Sinne verbesserter *User Experience* ein wesentliches Projektziel (*tool science*, Wolff 2015).

Die Entwicklung des spätlateinisch-altfranzösischen Tagsets wird im Projekt — auch mit Hilfe unseres modifizierten *InterText*-Tools — vorangetrieben. In unserem Posterbeitrag erläutern wir das Vorgehen und präsentieren erste Ergebnisse.

Fußnoten

1. http://www.palafra.org/
2. http://bfm.ens-lyon.fr/
3. http://universaldependencies.org/
4. http://sourceforge.net/projects/aligner/
5. http://www.statmt.org/moses/
6. https://bitbucket.org/rvwfels/paravoz2
7. http://wanthalf.saga.cz/intertext

Bibliographie

Burghardt, Manuel (2014): „Engineering annotation usability - Toward usability patterns for linguistic annotation tools“. Diss. Phil., Universität Regensburg, Institut für Information und Medien, Sprache und Kultur, urn:nbn:de:bvb:355-epub-307682.

Denooz, Joseph (1978): „L’ordinateur et le latin, techniques et methods“, in: *Revue de*

l'Organisation Internationale pour l'Etude des Langues Anciennes par Ordinateur 4.

Eger, Steffen / vor der Brück, Tim / Mehler, Alexander (2015): „Lexicon-assisted tagging and lemmatization in latin: A comparison of six taggers and two lemmatization methods", in: *LaTeCH 2015* 105.

Guillot, Céline / Prévost, Sophie / Lavrentiev, Alexei (2010): *Manuel de référence du jeu cattex09*. technical manual, UMR ICAR, CNRS/ENS-LSH. http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf

Neves, Mariana / Leser, Ulf (2014): „A survey on annotation tools for the biomedical literature", in: *Briefings in bioinformatics* 15 (2): 327–340.

Selig, Maria / Eufe, Rembert / Linzmeier, Laura (2015): *CoLaMer* (corpus du latin mérovingien). (im Erscheinen).

Vondricka, Pavel (2014): „Aligning parallel texts with intertext", in: *Proceedings of LREC 2014*.

Vogüé, Adalbert de / Antin, Paul (1979): *GREGOIRE LE GRAND, Dialogues* II. Cambridge University Press.

Von Foerster, Wendelin (1876): *Li Dialoge Gregoire lo Pape. Altfranzösische Uebersetzung des XII. Jahrhunderts der Dialogen des Papstes Gregor, mit dem lateinischen Original, einem Anhang: Sermo de Sapientia und Moralium in Iob Fragmenta, einer grammatischen Einleitung, erklärenden Anmerkungen und einem Glossar*, première partie: Textes. Paris: Champion.

Wolff, Christian (2015): „The case for teaching ‚tool science'. Taking software engineering and software engineering education beyond the confinements of traditional software development contexts", in: *Global Engineering Education Conference (EDUCON), 2015 IEEE* 932–938 10.1109/EDUCON.2015.7096085

Paraphrasenerkennung im Projekt *Digital Plato*

Kath, Roxana
roxana.kath@me.com
Universität Leipzig

Keilholz, Franz
franz.keilholz@tu-dresden.de
Technische Universität Dresden

Klinker, Fabian
fabian.klinker@tu-dresden.de
Technische Universität Dresden

Pöckelmann, Marcus
marcus.poeckelmann@informatik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg

Rücker, Michaela
mruecker1@me.com
Universität Leipzig

Švitek, Mihael
mihael.svitek@tu-dresden.de
Technische Universität Dresden

Wöckener-Gade, Eva
woeckener-gade@uni-leipzig.de
Universität Leipzig

Yu, Xiaozhou
xiaozhou.yu@tu-dresden.de
Technische Universität Dresden

Einleitung

Platons Werke wurden seit ihrer Entstehung bis in die heutige Zeit vielfach rezipiert und direkt zitiert, seine enorme Wirkungsmacht ist kaum zu unterschätzen, wie A.N. Whiteheads (1929, 63) berühmter Ausspruch verdeutlicht: " The safest general characterization of the European philosophical tradition is that it consists of a series of footnotes to Plato ". Mit dem von der VolkswagenStiftung geförderten Projekt *Digital Plato: Tradition and Reception* unter Leitung von Prof. Dr. Paul Molitor, Dr. Jörg Ritter, Prof. Dr. Joachim Scharloth, Prof. Dr. Charlotte Schubert und Prof. Dr. Kurt Sier wird seit April 2016 das Vorhaben verfolgt, diese Rezeption und Nachwirkung Platons bei den griechischen Autoren bis zur Spätantike systematisch zu untersuchen und zwar über das möglichst umfassende Auffinden von Paraphrasen. Der vorliegende Beitrag beschreibt die damit einhergehende grundlegende Problematik am Beispiel und skizziert einen derzeit in Entwicklung befindlichen Ansatz zu deren Lösung.

Die Textgrundlage

Die Werke Platons haben die Zeit weitestgehend überdauert und sind frei digital verfügbar [1]. In den Handschriften sind 43 Werke überliefert, mit Ausnahme der Apologie und der 13 Briefe alle in Dialogform verfasst. Heute werden die Dialoge in neun Gruppen zu je vier Schriften gruppiert (sogenannte Tetralogienordnung). Damit sind wir in der exzeptionellen Lage, alle Werke Platons untersuchen zu können, die in der Antike bekannt waren, auch wenn einige davon ihm fälschlicherweise zugeschrieben wurden oder die Autorenfrage umstritten ist. Diejenigen sieben Werke, die die sogenannte *Appendix Platonica* formen und schon in der Antike für nicht platonisch gehalten wurden, liegen vorerst nicht im Projektfokus (Erler 2006, 27-36).

Die Tetralogien haben einen Umfang von knapp 75.000 Zeilen. Dem gegenüber steht das wesentlich umfangreiche Gesamtwerk der antiken griechischen Autoren, das mit dem *Thesaurus Linguae Graecae* (TLG) in digitaler Form vorliegt und über neun Millionen Textzeilen umfasst.

Problemaufriss Paraphrasenerkennung

Um die Rezeption und Nachwirkung von Platons Werk in der antiken griechischen Literatur untersuchen zu können, sollen Übereinstimmungen zwischen seinen Texten und denen späterer Autoren im TLG gefunden werden. Dies geht bei weitem über das Identifizieren wörtlicher Zitate hinaus, da es das möglichst zuverlässige Auffinden paraphrasiert wiedergegebener Textstellen umfasst. Der Paraphrasenbegriff selbst wird im Rahmen des Projekts derzeit mit dem Arbeitskonzept der ‚Relation' bestimmt: Wie solche Relationen zwischen dem platonischen Werkkorpus und der übrigen griechischen Literatur der Antike aussehen und welche Aspekte damit erfasst werden können, soll folgendes Beispiel veranschaulichen:

Pl. symp. 206 d 1-2

ἀνάρμοστον δ᾽ ἐστὶ τὸ **αἰσχρὸν** παντὶ τῷ θείῳ, τὸ δὲ καλὸν ἁρμόττον.

Unvereinbar aber ist das Hässliche mit allem Göttlichen, aber das Schöne ist vereinbar.

Plot. enn. III 5, 1, 19-20

τὸ μὲν γὰρ **αἰσχρὸν** ἐναντίον καὶ τῇ φύσει καὶ τῷ θεῷ.

Denn das Hässliche ist sowohl der Natur als auch dem Gott entgegengesetzt.

Beim Paraphrasieren einer Textstelle kann es zu verschiedenen Phänomenen kommen. Neben fast wortwörtlicher Übernahme einer Textstelle (ggf. mit Auslassungen) können in den Textfluss eingewobene Zitate mit umgestelltem Satzbau auftreten. Das Beispiel geht darüber hinaus: Die wörtliche Übereinstimmung beschränkt sich auf einen geläufigen Ausdruck („das Hässliche"). Zudem wird der Inhalt einerseits nur teilweise wiedergegeben (die Vereinbarkeit vom „Schönen" und „Göttlichen" fehlt), andererseits um Neues erweitert (das „Hässliche" ist nun auch der „Natur" entgegengesetzt). Mit dem Synonym „ist entgegengesetzt" statt „ist unvereinbar" tritt eine lexikalische Varianz in Erscheinung. Zudem wurde das substantivierte Adjektiv „dem Göttlichen" samt seines attributiven Zusatzes „allem" durch das Nomen „dem Gott" ersetzt.

Ferner sind bspw. die Verwendung von Antonymen oder Metaphern denkbar, die die Erkennung einer Rezeption zusätzlich erschweren.

Vorarbeiten

Da sich die aufzufindenden paraphrasierten Textstellen nicht auf einen beliebigen Autor, sondern auf Platon beziehen, ergeben sich einige Vorteile für die Suche. Der überschaubare Umfang der Texte ermöglicht die manuelle bzw. teil-automatisierte Extraktion von Informationen aus den Werken Platons. Dazu gehören Listen mit den vorkommenden Substantiven, Verben, Eigennamen oder Stoppwörtern. Aber auch die Auflistung der zentralen Konzepte der platonischen Philosophie [2] ist für die spätere Paraphrasenerkennung hilfreich. Zudem liegen verschiedene Übersetzungen in elektronischer Form vor, die in das Projekt einfließen [3].

Einen großen Gewinn stellt auch die Vorarbeit des an der Universität Leipzig durchgeführten Projektes eAQUA [4] dar, welches ein Werkzeug zur Zitationsanalyse entwickelt hat und damit die vorkommenden Zitate im Korpus bereitstellt.

Für die Bewertung und Extraktion von Paraphrasen aus einem Textkorpus gibt es bereits verschiedene Ansätze, wie Androutsopoulos und Malakasiotis (2010) in einem Übersichtsartikel zusammengetragen haben. Diese basieren häufig auf einer Kontextanalyse und der Annahme, dass Worte in einem ähnlichen Kontext auch eine

ähnliche Bedeutung haben. So können für jedes Textsegment einer festen Länge (n-Gramme, meist mit n ≤ 5) die Kontexte aller Vorkommen betrachtet und als ein Vektor repräsentiert werden. Ähnlichkeitsmaße auf Vektoren erlauben nun den Vergleich zweier Textsegmente. Für das Auffinden von Paraphrasen müssen auf diesem Weg alle Textsegmente miteinander verglichen werden. Allerdings sind die so extrahierten Paraphrasen bzw. -fragmente sehr kurz, im Gegensatz zu den teils umfangreichen Rezeptionen, die im Rahmen des Projekts gefunden werden sollen. Zielführender sind Vorgehen, die zunächst Anker, d.h. eine Gemeinsamkeit zwischen zwei Textstellen, suchen und in einem zweiten Schritt die Fundstellen ausweiten. Solche Anker können bspw. einzelne Wörter, die oben beschriebenen n-Gramme sowie syntaktische oder semantische Repräsentationen einer Textstelle sein. Naheliegend ist, die Fundstellen im zweiten Schritt auf den umliegenden Satz auszuweiten. Stattdessen kann auch die Sinneinheit über semantische Informationen rekonstruiert werden, wie ein erfolgreich auf englischsprachige Korpora angewandtes Verfahren von Regneri, Wang und Pinkal (2014) aufzeigt.

Viele der bestehenden Verfahren zum Extrahieren von Paraphrasen erlangen ihre Effektivität mittels umfangreicher Annotationen der zu Grunde liegenden Texte, welche durch Parser mit einem gewissen Wirkungsgrad automatisch bestimmt werden können. Dieser Wirkungsgrad ist wiederum stark von zu Grunde liegenden Trainingskorpora und damit der Sprache der betrachteten Texte abhängig. So ist die Entwicklung für moderne Sprachen sehr weit vorangeschritten. Die Anwendung auf Altgriechisch ist hingegen deutlich seltener und auf weniger umfangreiche Korpora beschränkt (Mambrini und Passarotti 2012). Einer der ersten Vertreter ist das regelbasierte Analysewerkzeug Morpheus, das unter anderem Lemmata bestimmen kann (Crane 1991). In einer aktuellen Studie von Celano, Crane und Majidi (2016) wurden fünf aktuelle POS-Tagger mit Hilfe der Ancient Greek Dependency Treebank (Bamman und Crane 2011) trainiert und auf ihre Wirksamkeit getestet, wobei der Mate-Tagger [5] mit einer Genauigkeit von 88% am besten abschnitt. Das entsprechende Modell wurde dem Projekt zur Verfügung gestellt. Auch wenn die Parser sich stets weiterentwickeln, bleiben insbesondere die lange Zeitspanne und die vielfältigen Genres in dem von uns betrachteten Korpus problematisch, sodass die Parser und damit auch die darauf aufbauenden Verfahren zur Paraphrasenerkennung qualitativ schlechtere Ergebnisse produzieren als für moderne Sprachen (Dik und Whaling 2008).

Umsetzung im Projekt

Das verbreitete Vorgehen zur Extraktion von Paraphrasen über die Suche von Ankern wird für dieses Projekt durch die in Abschnitt 4 beschriebenen Vorarbeiten und die Entwicklung einer interaktiven Arbeitsumgebung für Suche und Auswertung von Paraphrasen praktikabel. Zwei Anwendungsszenarien sind dabei zu unterscheiden: die Suche ausgehend von einem Textstück und das Auffinden möglichst aller Rezeptionen Platons im Korpus. Der Fokus der aktuellen Arbeiten liegt zunächst in der ersten Aufgabe, ist ihre Bewältigung doch Grundlage für die zweite. Im Folgenden wird ein erster Ansatz beschrieben, der derzeit umgesetzt wird.

Ausgehend von einem Textstück, wie einem Satz von Platon, werden geeignete Anker für die Suche gewählt. Statt dabei alle Wörter zu berücksichtigen, kann die Auswahl auf Basis der angefertigten Listen auf bestimmte Wortarten oder auf die Begriffe der platonischen Philosophie beschränkt werden. Diese erste Vorfilterung reduziert die Anzahl der Suchwörter auf möglichst aussichtsreiche Kandidaten, um die anschließende Auswertung handhabbar zu halten. Dennoch ist eine gewisse Unschärfe, d.h. die Erweiterung eines Suchwortes zu einer Menge verwandter Worte, sinnvoll, um eine ganze Reihe von möglichen Rezeptionen abzudecken. Das Suchwort wird dazu durch die Verknüpfung von Wortrelationen erweitert, bspw. um seine Synonyme sowie verschiedene Übersetzungen samt deren Synonyme, die wiederum ins Altgriechische zurückübersetzt werden (siehe Abbildung 1).

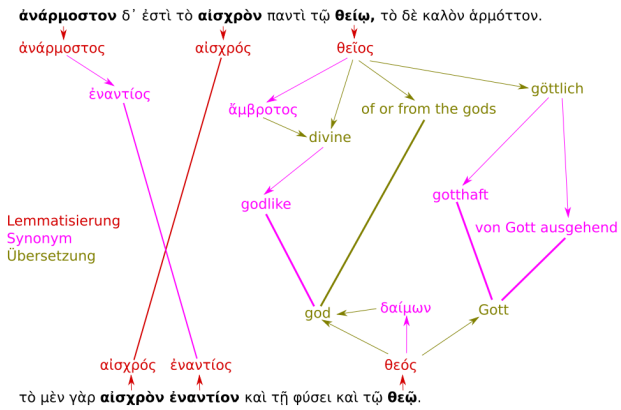


Abb. 1: Durch die Verknüpfung von Lemmatisierung, Synonymen und Übersetzungen kann das Suchwort θείῳ ('dem Göttlichen') um θεῷ ('der Gott') erweitert und so ein drittes Ankerpaar für das Beispiel gefunden werden.

Diese Erweiterung von Wortrelationen wird erfolgreich für die Verschlagwortung von Themen in Briefkorpora (Hildenbrandt et al. 2015) und in ähnlicher Form zum Aufbau eines *WordNet* für Altgriechisch (Bizzoni et al. 2014) genutzt. Ergänzt um zusätzliche Relationen, wie bspw. Hyperonym-respektive Hyponymbeziehungen, wird die Suche robuster gegenüber verschiedenen Formen der Paraphrasierung. Dabei gilt: Je mehr möglichst kurze Verbindungen zwischen zwei Wörtern liegen, desto größer ist die Aussagekraft dieses Paares. Wahrscheinliche Kandidaten für Rezeptionen sind dann Textstellen, in denen sehr viele aussagekräftige Anker nahe beieinander wiedergefunden werden.

Durch die Unschärfe des Verfahrens sind auch viele Kandidaten zu erwarten, die keine Paraphrasen sind und nicht als Rezeption des platonischen Werkes angesehen werden können. Die Ergebnisse sollen daher durch eine Arbeitsumgebung zunächst automatisch bewertet und sortiert werden. Ausgehend von einzelnen Treffern und einer transparenten Visualisierung, wie das System zur Entscheidung gelangte, soll eine interaktive Exploration der Texte die effiziente Recherche ermöglichen. Das beinhaltet das Wichten bzw. Entfernen einzelner Relationen sowie das manuelle Einordnen der gefunden Textstellen. So können Fallbeispiele und Phänomene näher untersucht, aber auch neue entdeckt werden. Die qualifizierte Bewertung auf der Basis der Fachexpertise von Altertumswissenschaftlern hilft wiederum, die Sammlung bereits bekannter Rezeptionen zu erweitern und die zu Grunde liegenden Algorithmen zu verbessern.

Die aus eAQUA bekannten Zitate sind für den Beginn des Projekts eine wichtige Unterstützung. Sie erlauben einen ersten Einblick in den Umfang der Rezeption Platons. Über die Verteilung lassen sich besonders häufig zitierte Passagen ermitteln, was möglicherweise auch Rückschlüsse auf die Fundstellen von Paraphrasen zulässt. Eine naheliegende, aber zu prüfende Hypothese ist, dass häufig zitierte Stellen auch anderweitig übernommen wurden. Das könnte zum zeitnahen Auffinden bisher unentdeckter Paraphrasen führen bzw. eine aufwendige Untersuchung an diesen Stellen rechtfertigen, um an besonders interessante Fallbeispiele zu gelangen.

Fußnoten

1. Siehe bspw. Perseus Digital Library http://www.perseus.tufts.edu/hopper/collection?collection=Perseus%3Acorpus%3Aperseus%2Cauthor%2CPlato
2. Eine entsprechende Liste findet sich bei Gigon und Zimmermann (1974, 301ff.)
3. Siehe bspw. Schleiermacher (Deutsch) oder Ü. Fowler (Englisch)
4. Siehe http://www.eaqua.net/
5. Siehe https://code.google.com/archive/p/mate-tools/

Bibliographie

Androutsopoulos, Ion / Malakasiotis, Prodromos (2010): „A Survey of Paraphrasing and Textual Entailment Methods“, in: *Journal of Artificial Intelligence Research* 38: 135–187.

Bamman, David / Crane, Gregory (2011): „Ancient Greek and Latin dependency treebanks“, in: *Language Technology for Cultural Heritage* 79–98 DOI:10.1007/978-3-642-20227-8_5.

Bizzoni, Yuri / Boschetti, Federico / Diakoff, Harry / Del Gratta, Riccardo / Monachini, Monica / Crane, Gregory (2014): „The Making of Ancient Greek WordNet“, in: *Proceedings of LREC 2010*.

Celano, Giuseppe G. A. / Crane, Gregory / Majidi, Saeed (2016): „Part of Speech Tagging for Ancient Greek“, in: *Open Linguistics* 2 (1), ISSN (Online) 2300–9969 10.1515/opli-2016-0020.

Crane, Gregory (1991): „Generating and Parsing Classical Greek“, in: *Literary and Linguistic Computing* 6 (4): 243–245 10.1093/llc/6.4.243

Dik, Helma / Whaling, Richard (2008): „Bootstrapping Classical Greek Morphology“, in: *DH2016: Book of Abstracts* 105–106.

Erler, Michael (2006): *Platon*. München: C.H.Beck.

Gigon, Olof / Zimmermann, Laila (1974): *Platon. Begriffslexikon*. Zürich: Artemis Verlag.

Hildenbrandt, Vera / Kamzelak, Roland S. / Molitor, Paul / Ritter, Jörg (2015): „im Zentrum eines Netzes [...] geistiger Fäden - Erschließung und Erforschung thematischer Zusammenhänge in heterogenen Briefkorpora“, in: *Datenbank-Spektrum : Zeitschrift für Datenbanktechnologie*: 15 (2015, 1): 49–55.

Mambrini, Francesco / Passarotti, Marco (2012): „Will a parser overtake Achilles? First experiments on parsing the Acient Greek Dependency Treebank“, in: *11th International*

Workshop on Treebanks and Linguistic Theories, Lisbon, Portugal.

Regneri, Michaela / Wang, Rui / Pinkal, Manfred (2014): „Aligning predicate-argument structures for paraphrase fragment extraction“, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Whitehead, Alfred North (1929): *Process and Reality: An Essay in Cosmology*. New York.

Raum und Zeit in Comics: Die Wirkung von Zwischenräumen auf Aufmerksamkeit und empfundene Zeit beim Lesen graphischer Literatur

Hohenstein, Sven
sven.hohenstein@uni-potsdam.de
Universität Potsdam, Deutschland

Laubrock, Jochen
laubrock@uni-potsdam.de
Universität Potsdam, Deutschland

Aufgrund der Kombination von Text und Bild stellen graphische Literatur und Comics komplexe Medien dar. Diese Hybridität stellt an die Aufmerksamkeit beim Lesen andere Anforderungen als bei rein textbasierten Romanen, da Informationen unterschiedlichen Typs erfasst und verarbeitet werden müssen. Wegen ihrer Konfiguration als eine Folge von Panels werden Comics auch als sequenzielle Kunst bezeichnet. Nach McCloud (1993) spielt der Raum zwischen den Panels, der als „gutter“ bezeichnet wird, eine Rolle für die Verbindung der einzelnen Panels. Obwohl dieser Raum selbst leer ist, so vergeht doch nach McCloud Zeit zwischen zwei Panels. Diesem Postulat hinsichtlich der Empfindung, die durch den „gutter“ ausgelöst wird, haben wir uns im Rahmen einer empirischen Studie gewidmet.

Die Wirkung zusätzlichen, leeren Raums zwischen Panels für die subjektive Wahrnehmung von Zeit beim Lesen graphischer Literatur haben wir mit kognitionspsychologischen Experimenten untersucht. Dieses Vorgehen erlaubt es über die reine Beschreibung des Materials hinaus den subjektiven Eindruck der Leserin bzw. des Lesers zu erfassen. Für diese Experimente stellten wir eine Sammlung von einzelnen Panels aus verschiedenen Comic-Reihen zusammen, beispielsweise „Astérix“ und „Donald Duck“. Die Auswahl der Panels erfolgte nach dem Kriterium, dass sie sich horizontal teilen lassen. Bei dieser Teilung wurde ein Panel per Bildbearbeitungssoftware in mehrere kleinere Unterpanels geteilt. Zusammenhängende Textabschnitte blieben dabei ungeteilt.

Im ersten Experiment wurde das Material in zwei Bedingungen dargeboten. In der Kontrollbedingung wurden die Panels jeweils ohne Teilung in ihrer ursprünglichen Form auf einem Bildschirm präsentiert. In der zweiten Bedingung wurden die Subpanels hintereinander auf dem Bildschirm gezeigt. Jeder Durchgang endete damit, dass die Probanden gefragt wurde, wieviel Zeit während der Geschichte, die in dem Panel erzählt wird, vergangen ist. Die Antworten der Probanden spiegeln somit deren subjektive Einschätzung der Dauer wider. Obwohl in beiden Bedingungen letztlich dieselben Panels gezeigt wurden, gab es bedeutsame Unterschiede in den Antworten. Die Teilung der Panels führte zu längeren subjektiven Dauern als die Kontrollbedingung. Dieses Ergebnis verdeutlicht den Einfluss der Konfiguration visueller Information auf die Wahrnehmung der Leserin bzw. des Lesers.

Um eine detailliertere Analyse der Aufmerksamkeit der Probanden vornehmen zu können, haben wir im zweiten Experiment zusätzlich Blickbewegungen erhoben. Für die Kontrolle der Auswirkungen der Panel-Teilung auf die wahrgenommene Dauer haben wir zudem das Material in einer Weise präsentiert, die ähnlicher zu tatsächlichen Comics ist. Die Subpanels wurden nebeneinander mit zusätzlichem, leerem Zwischenraum angeordnet, so dass das Aussehen einer kurzen Comic-Geschichte mit mehreren Panels gleicht. In der Kontrollbedingung wurden die Panels erneut ungeteilt dargeboten. Erneut wurden die Dauern länger eingeschätzt, wenn die Panels geteilt auf dem Bildschirm erschienen.

Die Auswertung der Blickbewegungen ergab ein differenziertes Bild der Aufmerksamkeitsverteilung beim Betrachten der Panels. Die Blickbewegungsmuster unterschieden sich in Hinblick auf die experimentelle Bedingung. Waren die Panels geteilt, so machten die Versuchspersonen mehr Fixationen. Die höhere Anzahl an Fixationen ist

somit eine mögliche Ursache für die subjektiv längere verstrichene Zeit. Außerdem zeigte sich eine leichte relative Tendenz zur Fixation nahe dem Zentrum eines jeden Subpanels, die bei geteilten Panels stärker ausgeprägt war. Diese und andere Befunde sprechen dafür, dass die Teilung von Panels die Aufmerksamkeit beim Lesen und Betrachten sowie die Wirkung graphischer Literatur beeinflussen kann.

Bibliographie

McCloud, Scott (1993): *Understanding comics: the invisible art*. Northampton, MA: Tundra.

relNet – Modellierung von Themen und Strukturen religiöser Online-Kommunikation

Elwert, Frederik
frederik.elwert@rub.de
RUB Bochum, Deutschland

Tabti, Samira
Samira.Tabti@ruhr-uni-bochum.de
RUB Bochum, Deutschland

Krech, Volkhard
volkhard.krech@rub.de
RUB Bochum, Deutschland

Morik, Katharina
katharina.morik@cs.uni-dortmund.de
Technische Universität Dortmund

Pfahler, Lukas
lukas@wandelt-pfahler.de
Technische Universität Dortmund

Von der Weiterentwicklung quantitativer Textanalysemethoden in der Informatik profitieren nicht nur die Geisteswissenschaften, auch für die qualitative Sozialforschung ergeben sich neue Impulse. Die Verwandtschaft qualitativer Forschungsmethoden in den Sozialwissenschaften und hermeneutischer Analyseansätze in den Geisteswissenschaften ermöglicht hier einen engen Austausch. Gleichzeitig stellt die Anwendung quantitativer Textanalysen zunehmend die ehemals strikte Trennung zwischen qualitativen und quantitativen Ansätzen in der Sozialforschung in Frage. Mit dem Aufkommen webbasierter Kommunikationsmedien können Sozialwissenschaftler_innen und Informatiker_innen zudem auf einen stetig wachsenden Datenbestand sozialer Interaktionen zugreifen, was zur Entstehung der *computational social sciences* als eigenem Forschungsfeld geführt hat (Lazer et al. 2009:721–23). Die Netzwerkanalyse hat sich hier zu einem der zentralen Methodenansätze entwickelt.

Die Religionswissenschaft ist ein dankbares Experimentierfeld für diese Art der disziplinen- und schulenübergreifenden Forschungsansätze. Sie vereint in sich sowohl geistes- als auch sozialwissenschaftliche Traditionen und ist in besonderer Weise am Zusammenspiel von Geistesgeschichte und sozialen Strukturen interessiert, wie dies bereits Max Weber mit seiner Unterscheidung von Ideen und Interessen (Weber 1989 [1920]: 101) herausgearbeitet hat. Daraus ergeben sich nach wie vor relevante Fragen: Wie prägen religiöse Vorstellungen das soziale Zusammenleben? Wie wirken sich aber auch soziale und politische Strukturen auf die Entwicklung und Weitergabe religiöser Ideen aus?

Das Projekt „relNet – Modellierung von Themen und Strukturen religiöser Online-Kommunikation“ nimmt vor diesem Hintergrund ein spezielles Segment gegenwärtiger Religiosität in den Blick: Neo-konservative christliche und islamische Bewegungen (etwa Evangelikale oder Anhänger der Salafiyya) haben in den letzten Jahren mit eigenen Online-Foren Kommunikationsplattformen geschaffen, in denen sie jeweils eigene Auslegungen in Theologie und Fragen der Lebensführung diskutieren (Becker 2009: 84; Neumaier 2016).

Bei allen Unterschieden zeichnen sich diese Bewegungen durch zwei Merkmale aus: a) eine Universalisierung von Religion im Sinne einer Ablösung „reiner“ Religion von Kultur und Politik, und b) eine religiöse Durchdringung aller Lebensbereiche, die sich insbesondere durch eine umfassende Regulierung der Lebensführung ausdrückt (O. Roy 2010: 25). Die Analyse dieser Online-Communities erlaubt es, Rückschlüsse über die Entwicklung und Verbreitung bestimmter Vorstellungen, aber auch über die Genese sozialer Strukturen und neuer Autoritäten zu ziehen.

Das Projekt ist eine Kooperation zwischen Religionswissenschaftler_innen der Ruhr-Universität Bochum und Informatiker_innen der TU Dortmund. Die interdisziplinäre Zusammenarbeit erlaubt es dabei insbesondere, Methoden in enger Passung auf das spezifische Material und die Fragestellungen zu adaptieren und zu entwickeln. In methodischer Hinsicht ist dabei die Unterscheidung von Strukturen und Inhalten leitend. Die Anwendung bereits etablierter Verfahren ermöglicht eine Analyse von Themen und ihrer zeitlichen Entwicklung einerseits (etwa über *topic models* wie LDA (Blei, Ng, and Jordan 2003)) sowie der sozialen Kommunikationsstrukturen in den Foren andererseits (etwa über *social network analysis*). Darüber hinaus werden im Rahmen des Projekts aber besonders solche Ansätze weiter erforscht, die beide Dimensionen in einem gemeinsamen Modell abbilden können. Dies bietet etwa die Möglichkeit, Gruppen in Netzwerken zu identifizieren, die sich nicht nur aufgrund ihrer Interaktionsstruktur, sondern auch durch gemeinsame Themen auszeichnen (Natarajan, N., Sen, P., & Chaoji, V. 2013: *2174–2177*).

Das Besondere dieses Ansatzes besteht darin, dass die vergleichsweise umfangreichen Datenbestände zu religiöser Onlinekommunikation nicht nur stichprobenartig, sondern in ihrer Gänze der Analyse zugänglich gemacht werden können

Die zeitliche Tiefendimension der Daten erlaubt zudem Analysen, welche die Themen nicht nur isoliert betrachten, sondern auch Diskursstränge in ihrer zeitlichen Abfolge und Verschränkung zu analysieren (Shahaf, Guestrin, and Horvitz 2012:1122–30).

Im Rahmen des Projekts werden dafür ausgewählte Online-Foren gecrawlt und in ein einheitliches Datenformat überführt. Für die eigentliche Analyse werden sie dann in die Software RapidMiner importiert, in der dann Verarbeitungs- und Analyseprozesse modelliert werden können. Neu entwickelte Methoden werden als Module für RapidMiner zur Verfügung gestellt. Dadurch lassen sich die Verarbeitungsschritte transparent dokumentieren und reproduzieren.

Das Poster stellt das Projekt sowie erste Zwischenergebnisse vor. Das Projekt wird vom Mercator Research Center Ruhr gefördert.

Bibliographie

Becker, Carmen (2009): „Gaining Knowledge: Salafi Activism in German and Dutch Online Forums“, in: *Masaryk University Journal of Law and Technology* 3 (1): 79–98 https://journals.muni.cz/mujlt/article/view/2526 [letzter Zugriff 15. November 2016].

Blei, David M. / Ng, Andrew Y. / Jordan, Michael I. (2003): „Latent Dirichlet Allocation“, in: *Journal of Machine Learning Research* 3: 993–1022 http://dl.acm.org/citation.cfm?id=944919.944937 [letzter Zugriff 15. November 2016].

Lazer, David / Pentland, Alex / Adamic, Lada / Aral, Sinan / Barabási, Albert-László / Brewer, Devon / Christakis, Nicholas et al. (2009): „Computational Social Science“, in: *Science* 323 (5915): 721–23 10.1126/science.1167742 .

Natarajan, Nagarajan / Sen, Prithviraj / Chaoji, Vineet (2013): „Community Detection in Content-Sharing Social Networks“, in: *In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* 496–500: 2174–2177.

Neumaier, Anna (2016): *Religion@home? Religionsbezogene Online-Plattformen Und Ihre Nutzung: Eine Untersuchung Zu Neuen Formen Gegenwärtiger Religiosität*. Religion in Der Gesellschaft 39. Ergon Verlag.

Roy, Olivier (2010): *Heilige Einfalt: Über Die Politischen Gefahren Entwurzelter Religionen*. München: Siedler.

Shahaf, Dafna, Carlos Guestrin, and Eric Horvitz (2012): „Metro Maps of Science“, in: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '12. New York, NY, USA: ACM 1122–1130 10.1145/2339530.2339706 .

Weber, Max (1989 [1920]): „Die Wirtschaftsethik Der Weltreligionen: Vergleichende Religionssoziologische Versuche: Einleitung“, in: *Max Weber Gesamtausgabe* 19. Tübingen: Mohr Siebeck 83–127.

„Soziale Datenkuratierung“: Nachhaltigkeit im Projekt *Illuminierte Urkunden als Gesamtkunstwerk*

Bürgermeister, Martina
martina.buergermeister@uni-graz.at
ZIM, Universität Graz, Österreich

Vogeler, Georg
georg.vogeler@uni-graz.at
ZIM, Universität Graz, Österreich

Das vom österreichischen Wissenschaftsfonds FWF geförderte Projekt hat sich zum Ziel gesetzt, die illuminierten Urkunden des Mittelalters zu sammeln, auf der Plattform *monasterium.net* zur Verfügung zu stellen und umfassend zu untersuchen. Die ExpertInnen aus den Bereichen der Diplomatik (Zajic Andreas, Gneiß Markus), der Kunstgeschichte (Roland Martin, Bartz Gabriele) und den Digitalen Geisteswissenschaften (Vogeler Georg, Bürgermeister Martina) arbeiten bewusst interdisziplinär zusammen und achten dabei vor allem auch auf Nachhaltigkeit. Das Poster wird zeigen, wie bei Materialerfassung, Erschließung und wissenschaftlicher Auswertung zukünftige Nutzerszenarien bedacht werden – also die soziale Dimension von Nachhaltigkeit konsequent berücksichtigt wird. Es reicht nicht, die Daten von Festplatte zu Festplatte zu kopieren (Langzeitarchivierung), sondern sie müssen auch nutzbar bleiben. Die im Projekt eingesetzten Mittel dafür sind 1. Datenkuratierung über eine etablierte Online-Plattform mit projektübergreifendem institutionellem Interesse, 2. Datenmanagement durch die Verwendung von gut dokumentierten und öffentlichen Datenstandards und 3. kontrollierte Vokabularien für die inhaltliche Erschließung.

Datenkuratierung

Schon in der Planungsphase des Projektes war klar, dass alle Projektdaten im weltweit größten Onlineangebot von Urkunden *monasterium.net* verarbeitet werden sollen. Damit wird ein sozialer Aspekt von Nachhaltigkeit berücksichtigt: Monasterium ist ein seit 2002 existierendes Großprojekt zur Zurverfügungstellung und (kollaborativen) Erschließung von Beschreibungen und Faksimiles von Urkunden des Mittelalters und der Frühen Neuzeit. Das Portal wird überwiegend von Archiven gespeist, es sind aber auch retrodigitalisierte Urkundenbücher und von ForscherInnen erstellte Sammlungen enthalten. Hinter dem virtuellen Archiv Monasterium steht *ICARUS*, ein Konsortium von Archiven und wissenschaftlichen Institutionen, das sein Wissen und seine Erfahrungen ständig austauscht und erweitert. Die große Datenmenge, die Etabliertheit des Angebots in der Fachcommunity und der institutionelle Hintergrund haben Monasterium aus einem kleinen DH-Projekt zu einem nachhaltigen Host nicht nur für unser Projekt gemacht: 1. Die projektübergreifende Infrastruktur erlaubt, dass die projektspezifischen Forschungsdaten über die Projektdauer hinaus zur Verfügung stehen. 2. Durch die Integration der Forschungsdaten in Monasterium bekommt jeder Datensatz auch einen persistenten Identifikator. D.h. alle Datensätze sind eindeutig adressierbar und zitierbar. 3. Das Interesse am Erhalt des Angebots ist groß, sodass selbst unter widrigen finanziellen Bedingungen aktiv nach Lösungen für den Erhalt der auf *monasterium.net* verfügbaren Daten gesucht werden wird.

Datenmanagement

Die im Projekt *Illuminierte Urkunden* entstehenden Forschungsdaten werden als strukturierte Datensätze in einer XML-Datenbank verwaltet und archiviert. Die einzelnen Urkunden-Datensätze werden nach dem Standard der CEI annotiert, die sich als TEI-P4-Dialekt in andere Datenstrukturen integriert und öffentlich dokumentiert ist. In der Datenbank ist ein für *monasterium.net* spezialisiertes Schema (XSD 1.1) im Einsatz, das einerseits die Verwendung der zulässigen Beschreibungselemente dokumentiert und andererseits die Konsistenz und Validität der zu importierenden Daten prüft. Schon in der Projektplanungsphase haben die Projektbeteiligten über Mittel und Möglichkeiten der Datenmodellierung gemeinsam diskutiert. Da der CEI-Standard zur wissenschaftlichen Bearbeitung von Urkunden initiiert wurde, brauchte die Überführung der aus dem Projekt *Illuminierte Urkunden* stammenden Forschungsdaten aus dem Bereich der Diplomatik keine Anpassungen an das Datenmodell. Um aber die Beschreibungsdaten zu Dekor und Buchschmuck aufnehmen zu können, musste das Datenmodell um die Möglichkeit einer kunsthistorischen Beschreibung erweitert werden. Dafür konnten Strukturen aus der TEI direkt übernommen werden. Die Daten werden also sozial nachhaltig, indem sie öffentlich dokumentierte und in der Fachcommunity geläufige Datenbeschreibungsstandards verwenden, Standards, die jede und jeder nachlesen kann.

Kontrollierte Vokabularien

Im Projekt *Illuminierte Urkunden* ist die Vergleichbarkeit von Datensätzen für die Weiternutzung ein wichtiger Faktor. Kulturelle Kontexte und Fragen der Mehrsprachigkeit spielen seit Projektstart eine Schlüsselrolle, da von Beginn an mit Forschungspartnern aus West-, Süd- und Südosteuropa zusammengearbeitet wird.

Bisher werden noch keine vollständigen Beschreibungen in mehreren Sprachen auf *monasterium.net* angeboten, aber im Rahmen des Projekts wurde die Möglichkeit entwickelt, Metadaten in mehrsprachigen kontrollierten Vokabularien zu erfassen. Sie werden im W3C SKOS als RDF/XML ausgedrückt. Bisher wurden ein viersprachiges kontrolliertes Vokabular zur Klassifikation des Dekors von Urkunden (vgl. Roland 2014) und ein Glossar erstellt. Diese Neuerung steigert die Qualität einerseits des Information Retrieval und führt zu einer umfassenden Kontextualisierung des Forschungsgegenstandes. Damit sind also auch die Inhalte der Daten für eine breitere Community besser nachvollziehbar, ein Konzept, das wir vorläufig als „Langzeitverständlichkeit“ bezeichnen möchten.

Zusammenfassung

Die kurze Laufzeit jedes Drittmittelprojektes – und damit auch des Projektes *Illuminierte Urkunden* macht Nachhaltigkeit als soziales Phänomen zu einer zentralen Frage: Die Forschungsdaten sollen einer sekundären Nutzung zur Verfügung gestellt werden und zu neuen Forschungsfragen führen. Deshalb werden im Projekt *Illuminierte Urkunden* drei „soziale“ Nachhaltigkeitsstrategien angewandt. Integration in eine in der Forschercommunity und bei Institutionen etablierte Plattform (monasterium.net), Verwendung von verbreiteten und facheinschlägigen Metadatenstandards und Erschließung von Inhalten mit kontrollierten Vokabularien.

Bibliographie

CEI, Charter Encoding Initiative: http://www.cei.lmu.de [letzter Zugriff 23. August 2016].

Heinz, Karl (2010): „Monasterium.net. Auf dem Weg zu einem europäischen Urkundeportal“, in: Kölzer, Theo (ed.): *Regionale Urkundenbücher*. Die Vorträge der 12. Tagung der Commission Internationale de Diplomatique, St. Pölten 2010 (Mitteilungen aus dem Niederösterreichischen Landesarchiv 14) 139–145.

ICARUS, International Centre for Archival Research: http://icar-us.eu/ [letzter Zugriff 23. August 2016].

Krah, Adelheid (2009): „Monasterium.net - das virtuelle Urkundenarchiv Europas: Möglichkeiten der Bereitstellung und Erschließung von Urkundenbeständen“, in: *AZ* 91: 221–246.

Roland, Martin (2013): „Illuminierte Urkunden im digitalen Zeitalter – Maßregeln und Chancen“, in: Ambrosio, Antonella / Barret, Sébastien / Vogeler, Georg (eds.): *Digital diplomatics. The computer as a tool for the diplomatist?*, Archiv für Diplomatik, Beiheft 14. Köln / Weimar /Wien 245–269.

Roland, Martin / Zajic, Andreas (2013): „Illuminierte Urkunden des Mittelalters in Mitteleuropa“, in: *Archiv für Diplomatik* 58: 237-428.

SKOS, Simple Knowledge Organization System: https://www.w3.org/2004/02/skos/ [letzter Zugriff 23. August 2016].

TEASys (Tübingen Explanatory Annotations System): Die erklärende Annotation literarischer Texte in den Digital Humanities

Zirker, Angelika
angelika.zirker@uni-tuebingen.de
Eberhard Karls Universität Tübingen, Deutschland

Bauer, Matthias
m.bauer@uni-tuebingen.de
Eberhard Karls Universität Tübingen, Deutschland

Das Poster präsentiert das Lehr- und Forschungsprojekt TEASys (Tübingen Explanatory Annotations System) zur erklärenden Annotation literarischer Text in den Digital Humanities. Die erklärende Annotation wird dabei als Anreicherung bislang

vor allem literarischer Texte um Informationen verstanden, die zum Textverständnis beitragen bzw. es überhaupt ermöglichen, d.h. sie dienen etwa der Überwindung von historischer Distanz (vgl. Hanna 1991). Eine Anwendung des Systems auf andere (nicht-literarische) Texte wird derzeit vorbereitet.

TEASys arbeitet mit verschiedenen Kategorien der erklärenden Annotation sowie ihrer Präsentation auf mehreren Ebenen, die sich etwa bezüglich ihrer Komplexität unterscheiden und aufeinander aufbauen (vgl. Bauer & Zirker 2015). Die Kategorien der erklärenden Annotation sind Sprache, Form, Intratextualität, Intertextualität, Kontext und Interpretation. Die Interpretation ergibt sich dabei aus den Informationen, die aus den anderen Kategorien zum besseren Verständnis an den Text herangetragen werden. Weitere Kategorien, die auf einer Meta-Ebene angesiedelt sind, beinhalten philologische Informationen (z.B. zu Varianten) sowie Fragen oder Anmerkungen (z.B. zu Items, zu denen bislang keine Informationen gefunden werden konnten sowie zur bislang bereits stattgefundenen Recherche zu einzelnen Items). Letztere Kategorie ist vor allem auch im Hinblick auf Fragen der Nachhaltigkeit essentiell. Die Ebenen der Annotation bauen aufeinander auf, d.h. die erste von insgesamt drei Ebenen bietet Informationen an, die das Textverstehen grundsätzlich ermöglichen, und die weiteren Ebenen nennen weitere, meist komplexere und ausführliche Informationen.

TEASys geht auf ein Peerlearning-Projekt zurück, das in Tübingen seit 2011 besteht und von Studierenden der englischen Literatur und weiteren geisteswissenschaftlichen Fächern getragen und von den Leitern des Forschungsprojekts (Prof. Dr. Matthias Bauer & PD Dr. Angelika Zirker) wissenschaftlich unterstützt wird. Es gibt derzeit vier Peerlearning-Gruppen, die sich mit Texten verschiedener Gattungen und Epochen beschäftigen und diese kollaborativ annotieren (zur Kollaboration in den DH s. z.B. McCarty 2012; Meister 2012; Stroud 2006). Das Forschungsprojekt widmet sich vor allem der Theoriebildung zur erklärenden Annotationen und der darauf aufbauenden Entwicklung eines best-practice-Modells, das wiederum auf die Theorie rückwirken soll (s. dazu Bauer & Zirker 2015). Die DH-Komponente liegt vor allem in der entsprechenden Aufbereitung und Visualisierung der erklärenden Annotationen für das digitale Medium sowie der darin möglichen Dynamik (s. Eggert 2009): Annotationen sind, entgegen ihrer Darstellung im Buch, ständig revidier- und erweiterbar und somit einer möglichst großen Rezipientengruppe offen, die umgekehrt für eine beständige Qualitätskontrolle sorgt. Ferner ermöglicht die digitale Repräsentation das Filtern von Informationen: je nach Bedarf können z.B. lediglich Annotationen zur Intertextualität angezeigt werden.

Das Poster stellt sowohl den Aufbau von TEASys als best-practice-Modell vor wie auch seine theoretischen Grundlagen und Beispielannotationen aus dem Peerlearning-Projekt, die von Studierenden erstellt wurden. Es macht deutlich, wie grundlegende hermeneutische Fragestellungen in das digitale Medium übernommen und dort abgebildet werden können (vgl. Drucker 2012) – und wie umgekehrt wiederum die digitale Präsentation aufgrund der theoretischen Überlegungen verbessert werden kann.

Bibliographie

Bauer, Matthias / Zirker, Angelika (2015): „Whipping Boys Explained: Literary Annotation and Digital Humanities“, in: Siemens, Ray / Price, Kenneth M: *Literary Studies in the Digital Age: An Evolving Anthology*. http://dlsanthology.commons.mla.org/under-review-matthias-bauer-and-angelika-zirker-whipping-boys-explained-literary-annotation-and-digital-humanities/ .

Drucker, Johanna (2012): „Humanistic Theory and Digital Scholarship“, in Gold, Matthew K. (ed.): *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press 85–95.

Eggert, Paul (2009): „The Book, the E-text and the ‚Work-site‘“, in: Deegan, Marilyn / Sutherland, Kathryn (eds.): *Text Editing, Print and the Digital World*. Ashgate 63–82.

Hanna, Ralph III (1991): „Annotation as Social Practice“, in: Barney, Stephan A. (ed.): *Annotation and Its Texts*. New York: OUP 178–184.

McCarty, Willard (2012): „Collaborative Research in the Digital Humanities“, in: Deegan, Marilyn / McCarthy, Willard (eds.): *Collaborative Research in the Digital Humanities*. Farnham: Ashgate 1–10.

Meister, Jan-Christoph (2012): „Crowd Sourcing ‚True Meaning‘: A Collaborative Approach to Textual Interpretation“, in: Deegan, Marilyn / McCarthy, Willard (eds.): *Collaborative Research in the Digital Humanities*. Farnham: Ashgate 105–122.

Stroud, Matthew D. (2006): „The Closest Reading: Creating Annotated Online Editions“, in: Bass, Laura R. / Greer, Margaret R. (eds.): *Approaches to Teaching Early Modern Spanish Drama*. New York: The MLA of America 214–219.

Tool zur Normalisierung und Historisierung

Eder, Elisabeth
e_eder@gmx.net
Ludwig Maximilians Universität München, Deutschland

Hadersbeck, Maximilian
maximilian@cis.uni-muenchen.de
Ludwig Maximilians Universität München, Deutschland

Das in diesem Poster vorgestellte, unter ‹http://goethefind.cis.uni-muenchen.de/?translator› verfügbare Translationstool überführt historisches Deutsch aus einem ungefähren Zeitraum von 1750 bis 1850 in gegenwartssprachliches Deutsch und umgekehrt modernen deutschen Text in seine historische Version.

Für eine Normalisierung oder Modernisierung von historischen Wörtern wurden in den letzten Jahren unterschiedliche Herangehensweisen präsentiert. Neben einer Modernisierung über Lexikon-Lookup, Transkriptionsregeln, Levenshtein-Distanz oder phonologische Ähnlichkeit fanden auch Methoden der statistischen maschinellen Übersetzung Anwendung (Scherrer / Erjavec 2015: 2f.). Um orthographischen Differenzen bei einer Translation einzelner Wörter aus eng verwandten Sprachen gerecht zu werden, werden dabei im Gegensatz zur standardmäßigen phrasenbasierten statistischen maschinellen Übersetzung die Phrasen nicht aus Wörtern, sondern aus Buchstabensequenzen gebildet und anstelle der Wörter der Ausgangs- und der Zielsprache die Buchstaben der Wortpaare aligniert [Pettersson et al., 2014]. Buchstabenbasierte statistische maschinelle Übersetzung zur Normalisierung historischer Wörter wurde vielfach mit dem Tool ‹Moses› [1] (Koehn et al. 2007) durchgeführt, wie beispielsweise bei (Pettersson et al. 2014), (Nakov / Tiedemann 2012) oder (Scherrer / Erjavec 2015). Neben einem Gebrauch zur Normalisierung wird dieses hier auch für die umgekehrte Überführungsrichtung eingesetzt.

In einem weiteren Ansatz zur Modernisierung und Historisierung bedient sich das Translationstool zudem neuronaler maschineller Übersetzung. Das dabei häufig verwendete Encoder-Decoder-Modell übertragen Faruqui, Tsvetkov, Neubig und Dyer (2016) auf die buchstabenbasierte Generierung von Wortflexion. Aufgrund der ähnlichen Grundlage kommt deren Tool ‹Morph-Trans› [2], das sich aus LSTMs, einer speziellen Form von rekurrenten neuronalen Netzen, zusammensetzt, zum Einsatz. Nach Wissen der Autoren ist dies der erste Versuch, ein neuronales Encoder-Decoder-Modell für eine Historisierung und Normalisierung deutscher Texte zu gebrauchen.

Als Trainings- und Entwicklungsdatensätze für die beiden Methoden dienten Wörter von 200 literarischen Texten aus einem Zeitraum von 1749 bis 1850. Diese Wörter wurden mithilfe des ‹Cascaded Analysis Broker› [3] vom Deutschen Textarchiv normalisiert, um im Anschluss daran auf die derzeit gültige ‹s›-Schreibung aktualisiert zu werden. Aus den historischen und den modernen Schreibweisen der Wörter wurden das Grundkorpus sowie ein Lookup-Lexikon gebildet. Im Translationstool werden die beiden Ansätze zusätzlich auch in Kombination mit diesem Lexikon eingesetzt. Zu Vergleichszwecken sind diese vier unterschiedlichen Ausgaben des Weiteren um ein auf einfachen Überführungsregeln und regulären Ausdrücken basierendes Verfahren ergänzt. Die unterschiedlichen Herangehensweisen können online anhand eigener Beispiele gegenübergestellt werden.

Tests auf exemplarischen Datensätzen zeigten, dass buchstabenbasierte statistische maschinelle Übersetzung nicht nur für eine Modernisierung, sondern im Deutschen ebenso für eine Historisierung dienlich ist und auch das neuronale Encoder-Decoder-Modell im Hinblick auf beide Überführungsrichtungen nutzbringend eingesetzt werden kann, wobei das Normalisieren im Vergleich zum Historisieren, wie zu erwarten war, durchweg bessere Resultate erzielte, was wohl unter anderem der Fülle an orthographischen Varianten in historischen Texten geschuldet ist.

Im geisteswissenschaftlichen Kontext ist eine Modernisierung historischer Wörter oftmals für eine erfolgreiche Anwendung sprachtechnologischer Werkzeuge, wie zum Beispiel Part-of-Speech-Tagger, auf älteren Texten von Nöten, während eine Historisierung beispielsweise bei der Suche

auf historischem Text zu einer erheblichen Erleichterung beitragen könnte, indem der moderne Suchterm historisiert wird, da von Anwendern und Anwenderinnen nicht erwartet werden kann, dass sie um die alten Schreibweisen der Wörter wissen. Eine Verwendung von buchstabenbasierter statistischer maschineller Übersetzung und buchstabenbasierten neuronalen Encoder-Decoder-Modellen zur Normalisierung und Historisierung bezüglich solcher Aufgaben und ähnlichen Problemstellungen im Bereich der Geisteswissenschaften ist vorstellbar.

Fußnoten

1. Online verfügbar unter: http://www.statmt.org/moses/
2. Online verfügbar unter: https://github.com/mfaruqui/morph-trans
3. Online verfügbar unter: http://www.deutschestextarchiv.de/demo/cab/

Bibliographie

Faruqui, Manaal / Tsvetkov, Yulia / Neubig, Graham / Dyer, Chris (2016): „Morphological Inflection Generation Using Character Sequence to Sequence Learning“, in: *Proceedings of NAACL* http://arxiv.org/pdf/1512.06110.pdf [letzter Zugriff 1. August 2016].

Koehn, Philipp / Hoang, Hieu / Birch, Alexandra / Callison-Burch, Chris / Federico, Marcello / Bertoldi, Nicola / Cowan, Brooke / Shen, Wade / Moran, Christine / Zens, Richard / Dyer, Chris / Bojar, Ondrej / Constantin, Alexandra / Herbst, Evan (2007): „Moses: Open Source Toolkit for Statistical Machine Translation“, in: *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic.

Nakov, Preslav / Tiedemann, Jörg (2012): „Combining Word Level and Character-Level Models for Machine Translation Between Closely Related Languages“, in: *Proceedings of ACL-2012*

Pettersson, Eva / Megyesi, Beáta / Nivre, Joakim (2014): „A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text“, in: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014* 32–41.

Scherrer, Yves / Erjavec, Tomaž (2015): „Modernising historical Slovene words“, in: *Natural Language Engineering* http://archiveouverte.unige.ch/unige:82305 [letzter Zugriff 1. August 2016]

Twhistory mit autoChirp Social Media Tools für die Geschichtsvermittlung

Hermes, Jürgen
hermesj@uni-koeln.de
Universität zu Köln, Deutschland

Hoffmann, Moritz
kontakt@moritz-hoffmann.de
Freier Historiker

Eide, Øyvind
oeide@uni-koeln.de
Universität zu Köln, Deutschland

Geduldig, Alena
ageduldi@uni-koeln.de
Universität zu Köln, Deutschland

Schildkamp, Philip
philip@schildkamp.net
Universität zu Köln, Deutschland

Public History (vgl. einführend Zündorf 2010) ist im deutschsprachigen Raum ein noch junges Feld, die erste Professur wurde erst Ende 2012 in Heidelberg eingerichtet. Die Disziplin ist zurückzuführen auf die doppelte Erkenntnis, dass die Mehrheit der Fachstudierenden nicht in der Geschichtswissenschaft wird arbeiten können (und dementsprechend zielgerichtet in Vermittlungskompetenzen aller Art geschult werden muss) und dass die meisten HistorikerInnen sich zwar über mangelnde Aufmerksamkeit für ihr Fach nicht beklagen können, demgegenüber aber kaum wissenschaftlich valide Werkzeuge für den Umgang mit der Öffentlichkeit entwickelt wurden.

Paradoxerweise scheint die Public History trotz ihres modernen Selbstanspruchs den Fehler der herkömmlichen Geschichtswissenschaft zu wiederholen: Die Digitalisierung ihrer Arbeit bleibt weit hinter den technischen Möglichkeiten

zurück und beschränkt sich größtenteils auf die Erleichterungen einer erweiterten Schreibmaschine. Doch Öffentlichkeiten, die sie schon ihrem Namen nach im Blick hat, migrieren zusehends in den digitalen Raum der sozialen Netzwerke und sollten genau dort angesprochen werden.

Eine Möglichkeit, die digitale Teilöffentlichkeit zu erreichen, bietet das soziale Netzwerk Twitter. Seit ungefähr sechs Jahren werden dort historische Ereignisse in je maximal 140 Zeichen zeitgenau nacherzählt, was unter den Bezeichnungen „Re-Entweetment“ oder auch “Twhistory” bekannt geworden ist. Dieses Potential des Medium wurde bislang fast ausschließlich von Laien genutzt, so über die Accounts @TitanicRealTime und das MDR-Projekt @9Nov89live , das über einen Tag eine fiktive Geschichte des Mauerfalls zeichnete. In jüngerer Zeit wird es aber zunehmend auch von einer geringen Zahl von (Public) Historians aktiv angeboten, beispielsweise für @NRWHistory und das Zweitweltkriegsprojekt @DigitalPast , zu dem parallel das Sachbuch “Als der Krieg nach Hause kam” (Hoffmann 2015) veröffentlicht wurde. Wahrscheinlich besser als jede andere Medienform bietet Twhistory die Möglichkeit der Erzählung in Echtzeit als nicht-textlichem Inhalt, über den Geschichte lebendig gemacht und vorhandenes historisches Interesse (re-)aktiviert werden kann.

Insbesondere die Zeichenbegrenzung ist für das Re-Entweetment Chance und Risiko zugleich: die Einstiegsschwelle ist im Vergleich zu herkömmlichen Darreichungsformen (Buch, Museum) äußerst gering, zugleich besteht die Gefahr der Simplifizierung sowie der Falschdarstellung von Geschichte als Aneinanderkettung von Einzelereignissen. Trotz der mittlerweile international steigenden Projektzahl hat sich noch keine Best Practice ergeben, um diesen Risiken zu begegnen. Dadurch ist auch die Zahl der digitalen Tools für diesen Bereich noch sehr klein, die Liste der Desiderate an die Digital Humanities aber lang und äußerst divers. Beispielsweise sind für die Planung, die Sammlung, die Gesamtschau und die Quellenreferenzierung von Inhalten Datenbanken oder zumindest tabellarische Aufstellungen notwendig, für die noch keine Möglichkeit bestand, die aggregierten Inhalte auch automatisch mit der Twitter-Plattform zu verknüpfen.

Dies hat sich mit der Bereitstellung der Software autoChirp geändert, die an der Kölner Informationsverarbeitung entwickelt wurde, um die Umsetzung entsprechender Twhistory-Projekte zu unterstützen. Zum einen vereinfacht autoChirp die Arbeit für die ErstellerInnen von Twitter-Timelines historischer Ereignisse, indem es eine Schnittstelle zum automatischen Upload von tabellarischen Sammlungen unterschiedlichen Formats anbietet. Dabei können neben dem gewünschten Datum, der genauen Uhrzeit und den Tweet-Text auch Bilder und Geolocations für den Tweet angegeben werden (vgl. Abb. 1). Auch können ganze Gruppen von Tweets per Mausklick auf eine neue Referenzzeit geschedulet werden.

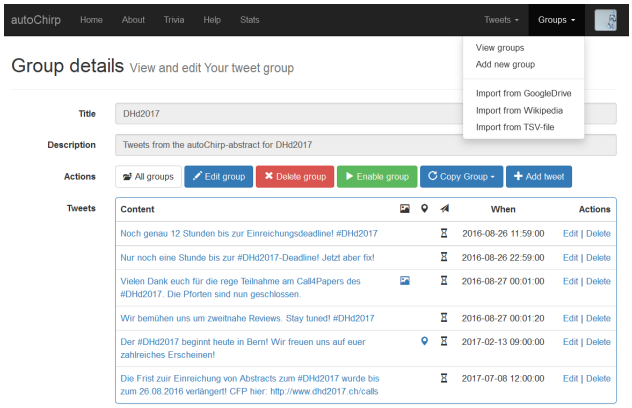


Abb. 1: Sceenshot des autoChirp-Web-Clients, mit dem eine Reihe von Tweets automatisch aus einer Tabelle geschedulet wurde. Das Web-Application-Frontend interagiert mit einer redundant angelegten Datenbank, um die Sicherung der in den verschiedenen Projekten generierten Tweets auch jenseits der Twitter-Plattform nachhaltig zu gewährleisten.

Die autoChirp-App wird zur Zeit mindestens von den Twitter-Projekten @DigitalPast (http://digitalpast.de/), @NRWHistory (http://nrwhistory.de/)und @goals_from_past genutzt und dabei unter anderem auch in der Lehre eingesetzt. Dabei stehen die EntwicklerInnen im engen Austausch mit den AnwenderInnen, um das Potential für Weiterentwicklungen abzuwägen. Aktuell wird die Integration von autoChirp in das Tiwoli-Projekt (vgl.Fischer & Strötgen 2015) realisiert, was zeigt, dass nicht nur historische, sondern auch literaturwissenschaftliche Vorhaben von einer Unterstützung im Zugang zur Twitter-Plattform profitieren können.
Für einen niederschwelligen Einstieg läuft eine Instanz von autoChirp als Web-Application zur freien Nutzung unter https://autochirp.spinfo.uni-koeln.de/ . Dort finden sich auch ausführliche Tutorials zur Benutzung. Für Weiterentwicklungen steht der dokumentierte Code im Github-Verzeichnis https://github.com/spinfo/autoChirp zur Verfügung.

Bibliographie

Fischer, Frank / Strötgen, Jannik (2015): „Wann findet die deutsche Literatur statt? Zur Untersuchung von Zeitausdrücken in großen Korpora“, in: *DHd 2015: Von Daten zu Erkenntnissen*.

Hoffmann, Moritz (2015): *Als der Krieg nach Hause kam*. Berlin: Ullstein.

Strötgen, Jannik / Gertz, Michael (2012): „Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards“, in *Proceedings of LREC 2012* 3746–3753.

Zündorf, Irmgard (2010): „Zeitgeschichte und Public History, Version: 1.0“, in: *Docupedia-Zeitgeschichte*, 11.2.2010 http://docupedia.de/docupedia/index.php?title=Public_History&oldid=68731 [letzter Zugriff 24. August 2016].

UIMA als Plattform für die nachhaltige Software-Entwicklung in den Digital Humanities

Hellrich, Johannes
johannes.hellrich@uni-jena.de
Graduiertenkolleg „Modell Romantik“, Friedrich-Schiller-Universität Jena, Jena, Deutschland

Matthies, Franz
franz.matthies@uni-jena.de
Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Deutschland

Hahn, Udo
udo.hahn@uni-jena.de
Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Deutschland

Texte und ihre automatische Analyse stehen im Zentrum vieler Untersuchungen in den Digital Humanities, etwa zur Erforschung sprachlicher und kultureller Wandlungsprozesse (siehe etwa Michel u.a. (2011)) oder im Bereich der Stilometrie (siehe etwa Jannidis (2014)). Die automatische Analyse von Texten beinhaltet typischerweise eine Reihe zunehmend komplexer werdender Schritte, angefangen bei der Segmentierung von Sätzen und Wörtern (Leerzeichen sind kein hinreichendes Kriterium, vgl. „*New York*“) über die syntaktische und semantische Analyse bis hin zu diskursstrukturellen und pragmatischen Analysen. Die für diese einzelnen Schritte nötigen sprachtechnologischen Komponenten sind oft, zumindest innerhalb einer Anwendungsdomäne, wiederverwendbar. Folglich gibt es mittlerweile eine Fülle von Software-Repositorien, die entsprechende computerlinguistische Komponenten sammeln, und Frameworks, die ihre Integration in sogenannte Pipelines, also funktionsbezogene sequenzielle Kombinationen von einzelnen Komponenten, erleichtern. Die dadurch ermöglichte Wiederverwendung von Komponenten ist im Sinne nachhaltiger Forschung, da diese so nicht mehrfach entwickelt werden müssen und der Software-Austausch zwischen Gruppen unterstützt wird.

Uima *(Unstructured Information Management Architecture)* [1] ist ein solches Framework, das sowohl im akademischen Kontext (in Deutschland u.a. DKPro [2] (de Castilho & Gurevych, 2014) und JCoRe [3] (Hahn u.a., 2016)) als auch in industriellen Anwendungen (etwa bei IBMs *Jeopardy* Champion Watson (Ferrucci u.a., 2010)) breite Verwendung findet (einen Vergleich unterschiedlicher Frameworks stellen Bank und Schierle (2012) an). Uima ist *open source* unter der Apache-Lizenz verfügbar und unterstützt mehrere Programmiersprachen, wobei Java in der Praxis eine dominierende Rolle zukommt.

Wir nutzen mit JCoRe seit fast einem Jahrzehnt Uima für computerlinguistische Problemstellungen in verschiedenen Domänen bzw. Sprachen und stellen die dabei entwickelten Komponenten öffentlich zur Verfügung. Aktuell arbeiten wir daran, unser ursprünglich für bio-medizinische Fragestellungen und englischsprachige Fachtexte entwickeltes Repositorium auf den DH-Bereich, primär für das Deutsche, zu erweitern. JCoRe stellt nicht nur sprachtechnologische Komponenten zur Verfügung, sondern auch die dafür nötigen Modelle für verschiedene Domänen — denn vor allem die Erstellung dieser Modelle ist ein enorm zeit- und rechenintensiver Prozess, der zudem ein hohes Maß an computerlinguistischer Expertise verlangt. Um die Einstiegshürden für die Benutzung solcher Ressourcen zu senken, bieten wir Anleitungen und Beispiele zur deklarativen Erstellung von Textanalyse-Pipelines mit Uima und haben zudem eine interaktive Anwendung entwickelt (Hahn u.a., 2016).

Eine Vielzahl von existierenden Sprachanalyse-Komponenten und Repositorien kann über Uima eingebunden werden, darunter auch einige, die nicht originär für das Framework entwickelt wurden, wie etwa das über DKPro verfügbare Stanford CoreNLP [4] (Manning u.a., 2014) oder OpenNLP [5]. Während Uima für den produktiven Einsatz entwickelt wurde, steht beim alternativen *Natural Language Toolkit* (NLTK) [6] der Einsatz in der Lehre im Zentrum (Bird u.a., 2009). Uima ist eher mit dem *General Architecture for Text Engineering* (GATE) Framework (Cunningham u.a., 2011) vergleichbar, das aber ein „geschlossenes" NLP-System repräsentiert, das exklusiv von den Entwicklern von Gate verwaltet wird. Generell sind integrierte Frameworks vorteilhaft gegenüber Pipelines aus einzelnen Werkzeugen, die mittels Textdateien/-strömen kommunizieren, da nicht bei jedem Schritt zwischen verschiedenen Formaten konvertiert werden muss. Insbesondere werden die bei selbstständigen Werkzeugen verbreiteten *in-line*-Annotationen (wie etwa *„das_Artikel Haus_Nomen"*) vermieden, die sich oft als unübersichtlich und fehleranfällig erweisen.

Uima und die anderen bisher genannten Frameworks sind primär für den Einsatz auf lokaler Rechner-Infrastruktur gedacht und somit nur bedingt mit Systemen wie WebLicht [7] (Hinrichs u.a., 2010) vergleichbar, die als Webservice verschiedene dezentral verteilte Komponenten zusammenführen. Dadurch wird zwar der Einstieg in die Nutzung sprachtechnologischer Systeme erleichtert, jedoch sind derartige Systeme nicht für die Verarbeitung großer Datenmengen geeignet und es entsteht eine eher intransparente Abhängigkeit von fremder Infrastruktur. Uima ist somit kein Konkurrent für WebLicht, sondern ermöglicht es vielmehr, Komponenten zu entwickeln, die bei Bedarf auch (durch in DKPro enthaltene Konverter) in WebLicht eingebunden werden können.

Im Kern ist Uima für die sequentielle Anreicherung mit Metadaten ausgelegt. Die möglichen Annotationen werden frei über ein objektorientiertes Typensystem definiert (siehe etwa Hahn u.a., 2007). In Uima wird zwischen Komponenten unterschieden, die Annotationen vornehmen *(Analysis Engines)*, und solchen, die Texte in das interne CAS *(Common Analysis System)* Format konvertieren *(Collection Reader)*; letztere können dabei auch bereits im Ursprungstext kodierte Metadaten verarbeiten. Die ersten Komponenten, die im Rahmen der Erweiterung JCoRes um DH-Komponenten entstanden und öffentlich zugänglich gemacht wurden, sind ein solcher *Collection Reader*, der die neuerdings vom *Deutschen Textarchiv* [8] (Geyken, 2013) zur Verfügung gestellten Dateien mit TCF- [9] und *Dublin Core*-Annotationen [10] verarbeiten kann, sowie eine entsprechende Erweiterung unseres Typensystems. In der unmittelbaren Zukunft geplante Erweiterungen betreffen *Analysis Engines* für Text- bzw. Wortsegmentierung und Wortartenerkennung (POS-Tagging) in historischen (literarischen) Texten.

Wir möchten durch unseren Beitrag insbesondere diejenigen, die primär computerlinguistische *Anwendungen* für Fragestellungen der Digital Humanities realisieren wollen (und damit meist keine computerlinguistischen *Entwicklung*sinteressen verfolgen), anregen, sich aus dem breiten Fundus existierender Komponenten zu bedienen und diese durch den Einsatz des Uima-Frameworks zu verbinden. Die dadurch implizit eingeführte Modularität erleichtert zudem die Durchführung von Funktionstests, die Anpassung an neue Domänen und darüber hinaus den Austausch mit anderen Forschenden — allesamt Anforderungen an eine nachhaltige Software-Infrastruktur.

Fußnoten

1. https://uima.apache.org
2. https:// DKPro .github.io
3. http://julielab.github.io
4. http://stanfordnlp.github.io/Core NLP
5. https://open NLP .apache.org
6. http://www.nltk.org
7. https://weblicht.sfs.uni-tuebingen.de
8. http://www.deutschestextarchiv.de/download
9. http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format
10. http://dublincore.org

Bibliographie

Bank, Mathias / Schierle, Martin (2012): „A survey of text mining architectures and the Uima standard", in: *Proceedings of LREC 2012* 3479–3486.

Bird, Steven / Klein, Ewan / Loper, Edward (2009): *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly.

de Castilho, Eckart R. / Gurevych, Iryna (2014): „A broad-coverage collection of portable

NLP components for building shareable analysis pipelines“, in: *OIAF4HLT 2014 – Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT @ COLING 2014* 1–11.

Cunningham, Hamish / Maynard, Diana / Bontcheva, Kalina (2011): *Text Processing with GATE*. Murphys, CA: Gateway Press.

Ferrucci, David A. / Brown, Eric / Chu-Carroll, Jennifer / Fan, James / Gondek, David C. / Kalyanpur, Aditya A. / Lally, Adam / Murdock, J. William / Nyberg 3rd, Eric H. / Prager, John M. / Schlaefer, Nico / Welty, Christopher A. (2010): „Building Watson: An overview of the DeepQA project“, in: *AI Magazine* 31 (3): 59–79.

Geyken, Alexander (2013): „Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv“, in: *Perspektiven einer corpusbasierten historischen Linguistik und Philologie* 221–234.

Hahn, Udo / Buyko, Ekaterina / Tomanek, Katrin / Piao, Scott / McNaught, John / Tsuruoka, Yoshimasa / Ananiadou, Sophia (2007): „An annotation type system for a data-driven NLP pipeline“, in: *LAW 2007 – Proceedings of the Linguistic Annotation Workshop @ ACL 2007* 33–40.

Hahn, Udo / Matthies, Franz / Faessler, Erik / Hellrich, Johannes (2016): „Uima-based JCoRe 2.0 goes GitHub and Maven Central: State-of-the-art software resource engineering and distribution of NLP pipelines“, in: *LREC 2016 – Proceedings of the 10th International Conference on Language Resources and Evaluation* 2502–2509.

Hinrichs, Erhard W. / Hinrichs, Marie / Zastrow, Thomas (2010): „WebLicht: Web-based LRT services for German“, in: *Proceedings of ACL-2010: System Demonstrations* 25–29.

Jannidis, Fotis (2014): „Der Autor ganz nah: Autorstil in Stilistik und Stilometrie“, in: Schaffrick, Matthias / Willand, Marcus (eds.): *Theorien und Praktiken der Autorschaft*. Berlin: de Gruyter 169–195.

Manning, Christopher D. / Surdeanu, Mihai / Bauer, John / Finkel, Jenny Rose / Bethard, Steven J. / McClosky, David (2014): "The Stanford CoreNLP Natural Language Processing Toolkit", in: *Proceedings of ACL-2014: System Demonstrations* 55–60.

Michel, Jean-Baptiste / Shen, Yuan K. / Aiden, Aviva P. / Veres, Adrian / Gray, Matthew K. / The Google Books Team / Pickett, Joseph P. / Hoiberg, Dale / Clancy, Dan / Norvig, Peter / Orwant, Jon / Pinker, Steven / Nowak, Martin A. / Aiden, Erez L. (2011): „Quantitative analysis of culture using millions of digitized books“, in: *Science* 331 (6014): 176–182.

Umfrage zu Forschungsdaten an der Philosophischen Fakultät der Universität zu Köln

Mathiak, Brigitte
bmathiak@uni-koeln.de
Universität zu Köln, Deutschland

Kronenwett, Simone
simone.kronenwett@uni-koeln.de
Universität zu Köln, Deutschland

Executive Summary

Um den aktuellen Bedarf an der Philosophischen Fakultät der Universität zu Köln im Umgang mit Forschungsdaten möglichst genau identifizieren zu können, wurde vom Data Center for the Humanities (DCH) in Kooperation mit dem Dekanat der Philosophischen Fakultät sowie der Universitäts- und Stadtbibliothek (USB) Köln 2016 eine Online-Umfrage unter dem akademischen Personal der Fakultät durchgeführt. Ziel der Erhebung ist es, sowohl die aktuellen Bestände zu charakterisieren, als auch Informationen zum Bedarf in den Bereichen Forschungsdatenmanagement (FDM) und Beratung zu erhalten. Im Vortrag werden die Ergebnisse der Umfrage präsentiert und diskutiert sowie mögliche Schlussfolgerungen erörtert.

Aktualität und Relevanz

Eines der wichtigsten neuen Handlungsfelder der Forschung, welche im Zuge der Digitalisierung von Information entstanden ist, betrifft das Management von Forschungsdaten. Die Hochschulen müssen sich darauf einstellen, ihren Wissenschaftlern und Forschern die notwendigen Infrastrukturen und Services zur Verfügung zu stellen. Auf diese Dringlichkeit verwies auch jüngst der Rat für Informationsinfrastrukturen in seinen Empfehlungen *Leistung aus Vielfalt* (Rat 2016). Denn noch immer gehen laut Schätzungen der

DFG bis zu 90% der digital produzierten Daten und Ergebnisse nach kurzer Zeit verloren bzw. "verschwinden in der Schublade" (Kramer 2014) und stehen somit keiner weiteren Verwendung und Nachnutzung zur Verfügung (Winkler-Nees 2011). Auch deshalb verabschiedete die Hochschulrektorenkonferenz (HRK) gleich zwei Grundsatzpapiere, in denen das Management von Forschungsdaten als zentrale strategische Herausforderung für die Hochschulleitungen angesehen wird (Hochschulschulrektorenkonferenz 2014 und 2015). Um einerseits die vielfältigen Aktivitäten und Akteure zu koordinieren und andererseits die Anschlussfähigkeit möglichst aller Hochschulen in den Scientific Communities auf nationaler und internationaler Ebene zu gewährleisten, erarbeitete die HRK einen 6-Punkte-Leitfaden, die sich aus ihrer Sicht beim Auf- oder Ausbau des institutionellen FDM ergeben und berücksichtigt werden sollen (Hochschulrektorenkonferenz 2015: 6-15). Im Rahmen dieses Maßnahmenkatalogs wird explizit empfohlen, zu Beginn eine Standortbestimmung an der jeweiligen Hochschule vorzunehmen, "z.B. mittels geeigneter interner Erhebungen zum Verhalten der Wissenschaftlerinnen und Wissenschaftler, aber auch zu deren Bedarfen." (Hochschulrektorenkonferenz 2015: 9)

Methodischer Ansatz

Der gewählte methodische Ansatz der Umfrage-basierten Studie orientiert sich an den sechs Leitlinien von (Müller et al. 2014) sowie an den einschlägigen Aufsätzen des Handbuchs *Methoden der Bibliotheks- und Informationswissenschaft* von Umlauf et al (Seadle 2013: 41-63; Fühles-Ubach 2013: 114-127; Fühles-Ubach 2013: 96-113; Fühles-Ubach, Umlauf 2013: 80-95).

Definition der Forschungsziele

Der Anlass für diese Studie ist die HRK-Empfehlung zur Durchführung einer Umfrage zu Forschungsdaten an Hochschulen als Grundlage für eine institutionelle FDM-Strategieentwicklung. Denn im Gegensatz zu einigen anderen deutschen Hochschulen fehlt für die drittgrößte Universität in Deutschland sowohl eine entsprechende Erhebung als auch weiterführend eine universitätsweite FDM-Policy. Der Fokus liegt allerdings *nicht* auf einer quantitativen Totalerhebung zu Forschungsdaten an der Kölner Volluniversität. Da professionelles FDM fachbereichsspezifisch erfolgen sollte (Sahle et al. 2013), richtet sich der Blick gezielt auf die Forschungsdaten an der Philosophischen Fakultät der Universität zu Köln, einer der größten geisteswissenschaftlichen Fakultäten Europas (UzK 2016). In diesem Kontext werden die Ergebnisse der Umfrage helfen, zur konzeptionellen Weiterentwicklung und Optimierung des DCH-Beratungs- und Serviceangebots, ein zentrales Dienstleistungsangebot der Fakultät, beizutragen.

Bestimmung der Zielgruppe

Die Teilnehmergruppe ist beschränkt auf das akademische Personal der Philosophischen Fakultät der Universität zu Köln. Die Umfrage zielte dabei besonders auf Wissenschaftler und Forscher, die direkt für datengestützte Forschungsprojekte verantwortlich sind.

Spezifizierung des Fragebogendesigns

Für die inhaltliche, konzeptionelle und methodische Gestaltung der Umfrage wurden die bisher verfügbaren Erhebungen zu Forschungsdaten an nationalen und internationalen wissenschaftlichen Institutionen und fachspezifischen Forschungseinrichtungen analysiert (forschungsdaten.org 2016; Burger et al. 2013) und auf die besonderen Gegebenheiten an der Philosophischen Fakultät zugeschnitten (Andorfer 2015; Stäcker 2015; CCeH 2016, DCH 2016a). Es wurden auch die Ergebnisse mehrerer Experteninterviews des DCH mit Wissenschaftlern der Philosophischen Fakultät berücksichtigt, die im Vorfeld der Erhebung im Rahmen von FDM-Beratungen durch das DCH geführt wurden. Der Fragebogen wurde insgesamt in fünf Teilbereiche untergliedert: 1) Forschungsdaten 2) Nutzung von Datenarchiven 3) Unterstützung beim Umgang mit Forschungsdaten 4) Fachbereich und Position 5) Interesse.

Überprüfung und Pretests

Eine Word-Version des Fragebogens wurde zunächst an alle Kooperationspartner des

Projektes verschickt. Nach Einarbeitung aller Rückmeldungen wurde der Fragebogen online programmiert und der Testlink an Wissenschaftler aller acht Fächergruppen der Fakultät sowie an externe Experten mit soziologischem Hintergrund für Pretests versendet. In mehreren Iterationen wurde der Fragebogen immer weiter adaptiert. Dies betraf u.a. die Reihenfolge der einzelnen Frageblöcke und die Auswahl der verwendeten Definitionen sowie die Präzision der Fragestellungen. Alle Schritte erfolgten in enger Absprache mit dem Datenschutzbeauftragten der Universität zu Köln.

Umsetzung und Einführung

Der Fragebogen wurde mit Hilfe des Online-Befragungstools von (Kronenwett&Adolphs 2016) erstellt und war vom 30.05.2016 bis 12.06.2016 aktiv. Die Auswertung der Ergebnisse erfolgte mit R.

Weiterführende Informationen zur Online-Umfrage sowie zu dem umfassenden Ergebnisbericht können auf der DCH-Webseite zur Umfrage eingesehen werden: http://dch.phil-fak.uni-koeln.de/umfrage-2016.html (DCH 2016b).

Deskriptive Datenanalyse: Ergebnisauswahl

Der Fragebogen wurde von 191 Personen begonnen und von 136 Teilnehmern vollständig beantwortet. 71.20% der Teilnehmer, die den Fragebogen begonnen haben, haben die Befragung auch beendet, d.h. sie haben alle Fragen vollständig beantwortet und sich bis zur Abschlussseite durchgeklickt. Die folgende Auswahl der Datenauswertung berücksichtigt nur diese Teilnehmer (n=136). Unser Ziel bei der Erstellung des Fragebogens war folgende Fragen zu beantworten:

- Welche Forschungsdaten gibt es?
- Welchen Bedarf gibt es bezüglich Forschungsdaten?
- Welche Unterstützung wünschen sich die Mitglieder der Fakultät von uns?

Zur ersten Frage war uns Nachhaltigkeit und Volumen wichtig. Zur Nachhaltigkeit konnten wir feststellen, dass die Mehrzahl der Befragten die Daten auf ihren lokalen Rechnern speichert: 70% auf dienstlichen Rechnern, 70% auf privaten Rechnern, Mehrfachantworten waren möglich (vgl. Abb. 1). Nur 14% speichern ihre Daten in einem Datenarchiv, eine Zahl, die sich auch in anderen Fragen reflektiert wird, etwa wie viele sich vorstellen können ihre Daten in einem Datenarchiv abzulegen.

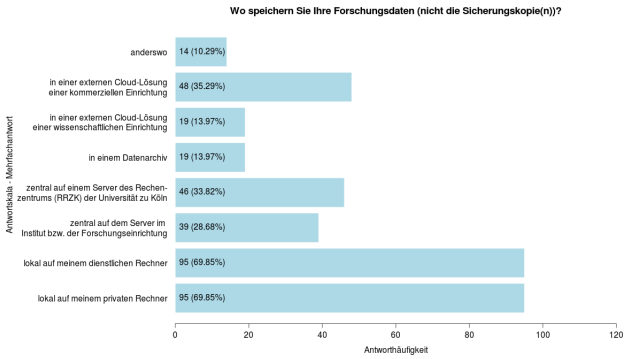


Abb. 1: Speicherort der Forschungsdaten (n=136)

Dies ist für die Nachhaltigkeit fatal, da nur in einem Datenarchiv ein strukturierter Zugriff und insbesondere auch Auffindbarkeit gewährleistet sind. Cloud-Lösungen, die auch weit verbreitet sind (35% Nutzung von kommerziellen Anbietern und 14% von wissenschaftlichen Anbietern), stellen zwar sicher, dass der Nutzer immerfort und von überall auf die Daten zugreifen kann und diese auch teilen kann. Aber für die Nachvollziehbarkeit von Forschungsergebnissen und die langfristige Sicherung sind diese denkbar ungeeignet.

Wir gehen davon aus, dass dies daran liegt, dass die Forscher ihre Handlungsweise nicht bezüglich Nachhaltigkeit und Nachvollziehbarkeit reflektieren. Die Selbsteinschätzung zu den eigenen Kenntnissen im Bereich FDM (vgl. Abb. 2) zeigt, dass die Kenntnisse größtenteils als durchschnittlich oder noch geringer (71%) eingeschätzt werden.

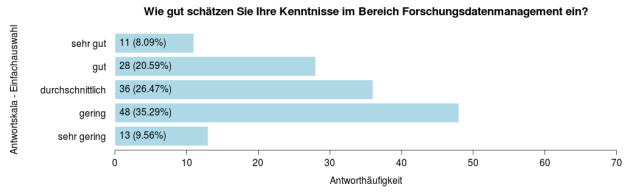


Abb. 2: Selbsteinschätzung der Kenntnisse im Forschungsdatenmanagement (n=136)

Es kann aber auch ein Faktor sein, dass selbst bei hohen Kenntnissen schlicht die Möglichkeiten fehlen, die Daten zu publizieren,

oder es keine Motivation bzw. Ressourcen gibt, dies auch tatsächlich zu tun.

Die Nachhaltigkeit an sich wird schon als Problem gesehen. 66% der Befragten geben an, dass sie befürchten die Daten zu verlieren, wenn sich nach Projektende niemand mehr für die dazugehörigen Webseiten zuständig fühlt. 60% fürchten Datenkonversionsprobleme. Aber auch für Probleme mit der Auffindbarkeit (45%) und der Dokumentation (41%) besteht eine prinzipielle Sensibilität. In Abbildung 3 finden Sie noch weitere Probleme, die von den Befragten genannt wurden. Interessant ist in diesem Zusammenhang auch, dass nur 11% der Befragten den Datenschutz bzw. die Datensicherheit als Problem sehen. Dies könnte aber auch im Zusammenhang damit stehen, dass wir explizit nach Problemen mit Forschungsdaten aus der Nutzerperspektive und nicht aus der Datengeberperspektive gefragt haben.

In unserer Beratungspraxis und auch in der Frage welche Serviceleistungen von einem Datenzentrum gewünscht werden (Abb. 4) spielen rechtliche Aspekte und Zugriffseinschränkungen eine sehr große Rolle. 74% wünschen sich diesbezüglich eine Beratung. Es ist damit das meistgewünschte Thema. Ebenfalls in der Spitzengruppe sind Beratungen zu technischen Themen (73%) und allgemeiner Natur (66%), sowie die konkrete Bereitstellung von Speicherplatz zur Archivierung und Publikation von Forschungsdaten (72%). Im Mittelfeld wird Unterstützung beim Erstellen eines Datenmanagementplans, z.B. für Drittmittelanträge gewünscht (54%), Beratung für Archivierung und Zitation (50%) und der Betrieb von laufenden Anwendungen (46%). Letzteres gestaltet sich für uns äußerst schwierig umzusetzen. Auch in der Beratungspraxis werden wir immer wieder mit diesem Wunsch konfrontiert. Die technischen Hürden und notwendigen Ressourcen sind jedoch zum Teil beträchtlich.

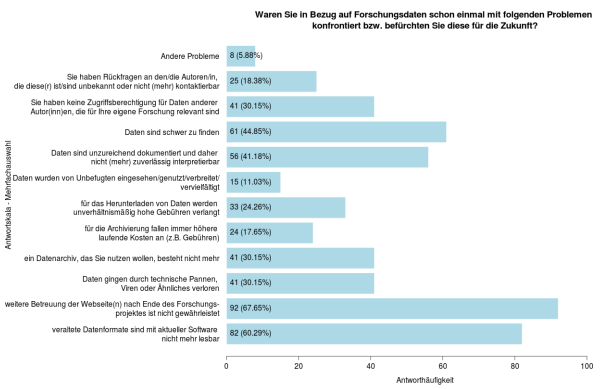


Abb. 3: Allgemeine Probleme mit Forschungsdaten (n=136)

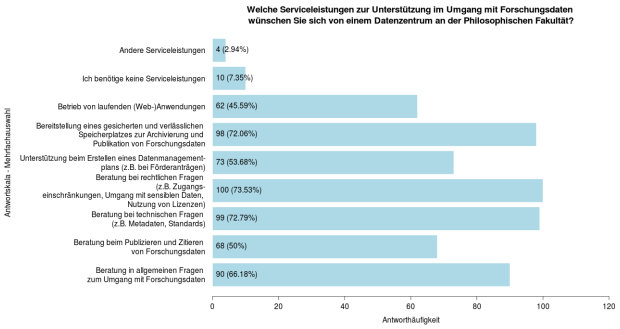


Abb. 4: Gewünschte Serviceleistungen (n=136)

Schlussfolgerungen und Ausblick

Im Vortrag werden wir noch genauer auf unsere Ergebnisse eingehen und diese auch mit anderen Studien vergleichen, die bereits an anderen Hochschulen sowohl in Deutschland (forschungsdaten.org 2016) als auch international (Kuipers et al. 2009; Bauer et al. 2016) durchgeführt wurden. Umfragen, wie die von uns durchgeführte, sind ein wichtiges Mittel für die strategische Positionierung von Institutionen, die sich mit Forschungsdaten beschäftigen. Wir werden daher im Vortrag auch kurz darauf eingehen, wie die von uns erhobenen Ergebnisse die Strategie des DCH (Data Center for the Humanities) an der Philosophischen Fakultät zu Köln beeinflusst hat (Kronenwett 2017).

Bibliographie

Andorfer, Peter (2015): *Forschung und Forschungsdaten in den Geisteswissenschaften: Zwischenbericht einer Interviewreihe*. DARIAH-DE working papers 10. Göttingen: GEODOC, Dokumenten- und Publikationsserver der Georg-August-Universität Göttingen http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2015-10.pdf [letzter Zugriff 30. August 2016].

Bauer, Bruno / Ferus, Andreas / Gorrais, Juan et al. (2015): *Forschende und ihre Daten. Ergebnisse einer österreichweiten Befragung*. Report 2015, Version 1.2 https://phaidra.univie.ac.at/view/o:407513 [letzter Zugriff 30. August 2016].

Burger, Marleen / Kindling, Maxi / Liebenau, Lisa et al. (2013):

Forschungsdatenmanagement an Hochschulen. Internationaler Überblick und Aspekte eines Konzepts für die Humboldt-Universität zu Berlin. Version 1.1 vom 3. März 2013 http://edoc.hu-berlin.de/oa/reports/reZ8xHXx2cLyc/PDF/28q8QGlHKwrRw.pdf [letzter Zugriff 30. August 2016].

Cologne Center for eHumanities (eds.) (2016): *Digital Humanities*. Strukturen - Lehre - Forschung. Universität zu Köln http://cceh.uni-koeln.de/broschure-digital-humanties-2016/ [letzter Zugriff 30. November 2016].

Data Center for the Humanities (2016a): Homepage, http://dch.phil-fak.uni-koeln.de/ [letzter Zugriff 12. August 2016].

Data Center for the Humanities (2016b): Homepage, Unterseite "Umfrage Forschungsdaten 2016", http://dch.phil-fak.uni-koeln.de/umfrage-2016.html [letzter Zugriff 12. August 2016].

Drees, Bastian (2016): „Zukunft der Informationsinfrastrukturen: Das deutsche Bibliothekswesen im digitalen Zeitalter“, in: *Perspektive Bibliothek* 5.1: 25–48 urn:nbn:de:bsz:16-pb-313858 .

DV-ISA (2016): *Umgang mit digitalen Daten in der Wissenschaft: Forschungsdatenmanagement in NRW. Eine erste Bestandsaufnahme*, 14. April 2016, Version 0.7 [Final] https://www.dh-nrw.de/fileadmin/dh-nrw/PDF/Veroeffentlichungen/DV-ISA-Bestandsaufnahme_FDM.pdf [letzter Zugriff 10. August 2016].

Fühles-Ubach, Simone (2013a): „Quantitative Befragungen“, in: Umlauf, Konrad / Fühles-Ubach, Simone / Seadle, Michael (eds.): *Handbuch Methoden der Bibliotheks- und Informationswissenschaft. Bibliotheks-, Benutzerforschung, Informationsanalyse*. Berlin: De Gruyter 96–113.

Fühles-Ubach, Simone (2013b): „Online-Befragungen“, in: Umlauf, Konrad / Fühles-Ubach, Simone / Seadle, Michael (eds.): *Handbuch Methoden der Bibliotheks- und Informationswissenschaft. Bibliotheks-, Benutzerforschung, Informationsanalyse*. Berlin: De Gruyter 114–127.

Fühles-Ubach, Simone / Umlauf, Konrad (2013): „Quantitative Methoden“, in: Umlauf, Konrad / Fühles-Ubach, Simone / Seadle, Michael (eds.): *Handbuch Methoden der Bibliotheks- und Informationswissenschaft. Bibliotheks-, Benutzerforschung, Informationsanalyse*. Berlin: De Gruyter 80–95.

forschungsdaten.org (2016): Homepage, Unterseite "Umfragen zum Umgang mit Forschungsdaten an wissenschaftlichen Institutionen", http://www.forschungsdaten.org/index.php/Umfragen_zum_Umgang_mit_Forschungsdaten_an_wissenschaftlichen_Institutionen [letzter Zugriff 29. August 2016].

Hochschulrektorenkonferenz (2014): *Management von Forschungsdaten als strategische Aufgabe der Hochschulleitungen*. Empfehlung der 16. HRK-Mitgliederversammlung am 13. Mai 2014 in Frankfurt am Main, https://www.hrk.de/uploads/tx_szconvention/HRK_Empfehlung_Forschungsdaten_13052014_01.pdf [letzter Zugriff 29. August 2016].

Hochschulrektorenkonferenz (2015): *Wie Hochschulleitungen die Entwicklung des Forschungsdatenmanagements steuern können. Orientierungspfade, Handlungsoptionen, Szenarien*. Empfehlung der 19. Mitgliederversammlung der HRK am 10. November 2015 in Kiel, https://www.hrk.de/uploads/tx_szconvention/Empfehlung_Forschungsdatenmanagement__final_Stand_11.11.2015.pdf [letzter Zugriff 29. August 2016].

Kramer, Bernd (2014): „Datenflut an Unis: Forscher müssen teilen lernen“, in: *Spiegel Online*, 26. Februar 2014, http://www.spiegel.de/unispiegel/jobundberuf/umgang-mit-daten-der-glaeserne-forscher-a-954958.html [letzter Zugriff 01. August 2016].

Kronenwett, Simone (2017): *Forschungsdaten an der Philosophischen Fakultät der Universität zu Köln* (= Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft 78). Technische Hochschule Köln.

Kronenwett&Adolphs (2016): Homepage, http://www.kronenwett-adolphs.com/de [letzter Zugriff 12. August 2016].

Müller, Hendrik / Sedley, Aaron / Ferrall-Nunge, Elizabeth (2014): „Survey research in HCI“, in: Olson, Judith S. / Kellogg, Wendy A. (eds.): *Ways of Knowing in HCI*. New York: Springer 229–266.

Philosophische Fakultät (2016): Homepage, Unterseite "Fächergruppen", http://phil-fak.uni-koeln.de/9785.html [letzter Zugriff 12. August 2016].

Rat für Informationsinfrastrukturen (2016): *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*. Göttingen, http://www.rfii.de/?wpdmdl=1998 [letzter Zugriff 30. August 2016].

Sahle, Patrick / Kronenwett, Simone (2013): „Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner ‚Data Center for the Humanities‘“, in: *LIBREAS. Library Ideas* 23: 76–96 urn:nbn:de:kobv:11-100212726.

Schöpfel, Joachim / Prost, Hélène (2016): „Research data management in social sciences and humanities: A survey at the University of Lille (France)“, in: *LIBREAS. Library Ideas* 29: 98–112 urn:nbn:de:kobv:11-100238193.

Seadle, Michael (2013): „Entwicklung von Forschungsdesigns“, in: Umlauf, Konrad / Fühles-Ubach, Simone / Seadle, Michael (eds.): *Handbuch Methoden der Bibliotheks- und Informationswissenschaft: Bibliotheks-, Benutzerforschung, Informationsanalyse*. Berlin: De Gruyter 41–63.

Stäcker, Thomas (2015): „Noch einmal: Was sind geisteswissenschaftliche Forschungsdaten?“, in: *DHdBlog*, 06.12.2015, http://dhd-blog.org/?p=5995 [letzter Zugriff 30. Juli 2016].

Universität zu Köln (2016): Homepage, Unterseite "Philosophische Fakultät", http://phil-fak.uni-koeln.de/studieninteressierte.html?&L=0 [letzter Zugriff 30. Juli 2016].

Winkler-Nees, Stefan (2011): „Vorwort“, in: Büttner, Stephan / Hobohm, Hans-Christoph / Müller, Lars (eds.): *Handbuch Forschungsdatenmanagement*. Bad Honnef: Bock +Herchen Verlag 5–6.

Visuelle Elemente grafischer Literatur: Aufmerksamkeitszuwendung und objektive Beschreibung

Laubrock, Jochen
laubrock@uni-potsdam.de
Universität Potsdam, Deutschland

Richter, Eike
Eike.Richter@uni-potsdam.de
Universität Potsdam, Deutschland

Hohenstein, Sven
hohenstein@uni-potsdam.de
Universität Potsdam, Deutschland

Graphische Romane vereinen als hybride Gattung Aspekte von Literatur und bildender Kunst (McCloud, 1993). Wie interagieren Bild und Text beim Lesen graphischer Literatur und ermöglichen das Verstehen des Gesamtwerkes? Worauf fokussiert die Aufmerksamkeit des Lesers? Als Methode zur Beantwortung dieser Fragen ist die Blickbewegungsmessung besonders geeignet. Blickbewegungen haben sich in einer Vielzahl an Studien als valides, nichtreaktives Maß für die Verarbeitung und das Verstehen von Text und Bild erwiesen, in dem sich zudem auch unbewusste Verarbeitungsprozesse niederschlagen (Findlay & Gilchrist, 2003; Wade & Tatler, 2005).

In früheren Arbeiten (Laubrock, Hohenstein & Thoß, 2016; Dunst, Hartel, Hohenstein Laubrock, 2016) haben wir mit Eyetracking-Analysen gezeigt, dass beim Lesen grafischer Literatur der größte Teil der Aufmerksamkeit dem Text in Sprechblasen und Beschriftungen (Captions) zugewandt wird und nur ein relativ kleiner Teil den originär visuellen Gestaltungselementen alloziert wird. Wird der visuelle Inhalt gar nicht beachtet, oder kann er möglicherweise bereits im peripheren Sehen während der Fixationen auf dem Text verarbeitet werden? Wir hatten bereits berichtet, dass Comics-Experten den Bildanteil stärker beachten und darauf verstehensrelevante Information extrahieren. In einer neuen Serie von Studien untersuchen wir mittels blickkontingenter Präsentation, ob (a) den Bildanteilen mehr Aufmerksamkeit zugewandt wird, wenn die Vorschau verhindert wird, indem das Bild erst eingeblendet wird, wenn der Blick sich auf ein Panel bewegt und (b) die Aufmerksamkeit andere grafische Elemente auswählt, wenn zwar der visuelle Teil der Panels sichtbar ist, der Text aber erst nach Fokussierung eines Panels eingeblendet wird.

Das visuelle Material wurde auf zweierlei Weise annotiert. Einerseits annotierten Menschen Personen und einzelne Objekte innerhalb der Panels. Andererseits versuchen wir eine objektiven Beschreibung des visuellen Materials mithilfe von Deskriptoren aus dem maschinellen Sehen (Computer Vision), z.B. mittels Farbhistogrammen, lokalem Fourier-Spektrum oder SIFT-Deskriptoren (Lowe, 1999). Der Vorteil dieser Beschreibung ist neben der Objektivität die skriptgesteuerte Anwendbarkeit auf große Datenmengen, etwa digitalisierte Korpora grafischer Literatur. Vergleichbare Arbeiten aus der Schnittstelle von Kunstgeschichte und Informatik ermöglichen beispielsweise eine automatisierte Klassifikation von Kunstrichtungen (Saleh & Elgammal, 2015) und zeigen das Potenzial eines solchen Ansatzes als Stilometrie visueller Merkmale.

Für die Zuordnung der Blickbewegungsdaten auf das Stimulusmaterial nutzen wir die im Projekt entwickelte Graphic Novel Markup

Language (GNML), eine Erweiterung der Comic Book Markup Language (CBML; Walsh, 2012). Das Material wurde mit unserem Editor annotiert, für Weiterverarbeitung und statistische Analyse der Daten nutzten wir ein in Entwicklung befindliches R-Paket. Die objektive Beschreibung des visuellen Materials mit Deskriptoren aus dem maschinellen Sehen wurde unter Nutzung von OpenCV (Bradski, 2000) und VLFEAT (Vedaldi & Fulkerson, 2008) teils in Python und teils in Matlab implementiert, da für R für diesen Anwendungsbereich keine hinreichend entwickelte Funktionsbibliothek existiert.

Bibliographie

Bradski, Gary (2000): „The OpenCV library“, in: *Dr. Dobb's Journal of Software Tools* 25 (11): 120–125.

Dunst, Alexander / Hartel, Rita / Hohenstein, Sven / Laubrock, Jochen (2016): „Corpus Analyses of Multimodal Narrative: The Example of Graphic Novels“, in: *DH2016: Conference Abstracts* 178–180.

Findlay, John M. / Gilchrist, Ian D. (2003): *Active Vision. The Psychology of Looking and Seeing*. Oxford: Oxford University Press.

Laubrock, Jochen / Hohenstein, Sven / Thoß, Aalexander (2016): „Moving around the city of glass“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 186.

Lowe, David G. (1999): „Object recognition from local scale-invariant features“, in: *Proceedings of the International Conference on Computer Vision (ICCV'99)* 1150–1157.

McCloud, Scott (1993): *Understanding comics: the invisible art*. Northampton, MA: Tundra.

Saleh, Babak / Elgammal, Ahmed M. (2015): „Large-scale classification of fine-art paintings: Learning the right metric on the right feature“, in: *CoRR* abs/1505.00855, 1–21 http://arxiv.org/pdf/1505.00855v1.pdf .

Vedaldi, Andrea / Fulkerson, Brian (2008): *VLFeat: An open and portable library of computer vision algorithms*. [Computer Software: http://www.vlfeat.org/]

Wade, Nicholas J. / Tatler, Benjamin W. (2005): *The Moving Tablet of the Eye: Origins of modern eye movement research*. Oxford: Oxford University Press.

Walsh, John (2012): „Comic Book Markup Language: An Introduction and Rationale“, in: *DHQ: Digital Humanities Quarterly* 6 (1).

… warum nicht gleich Wikidata?!

Schelbert, Georg
georg.schelbert@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Das digitale Bildformat verleiht analogen Bildsammlungen eine zweite Existenz. Insbesondere aber schafft die Verbindung des digitalen Formats mit dem Internet einen weitgehend raumunabhängigen universellen Bilderpool, der die kaum fassbare Menge der Produktion überhaupt erst sichtbar werden lässt.

Die Mediathek des Instituts für Kunst- und Bildgeschichte der HU besitzt eine umfangreiche Sammlung historischer Fotografien und Diapositive, die bislang weitgehend unerschlossen ist. Diese wissenschaftliche Lehrsammlung eines traditionsreichen kunsthistorischen Universitätsinstituts, bestehend aus zwischen ca. 1890 und 1970 hergestellten Glasdias im Format 8,5 x 10cm ist auch hinsichtlich ihres Umfangs von etwa 60.000 Stück herausragend.

Einerseits spiegelt die Sammlung Interessensschwerpunkte großer, an der Berliner Universität lehrender Fachgelehrter wie Heinrich Wölfflin, Adolph Goldschmidt, Wilhelm Pinder und Richard Hamann wider, andererseits repräsentiert sie den gesamten, an der Universität über Jahrzehnte hinweg geformten Kanon der Kunstgeschichte, der inzwischen auch zu allgemeinem Bildungsgut geworden ist.

Damit ist diese Sammlung ein typisches Beispiel eines kunsthistorischen Bildbestandes, der sich dadurch auszeichnet, dass er vor allem Repräsentationen von Werken beinhaltet, die als solche bereits vielfach identifiziert und erschlossen sind. Aus diesem Grund besteht die Aufgabe zunächst darin, die Bilddatei mit bereits vorhandenen Wissensbeständen zu verbinden. Auch die vielen jüngeren Ansätze zum Umgang mit digitalen Bildern - Automatische Bilderkennung, Folksonomy-Tagging, Festlegung von Metadatenstandards einschließlich der Verwendung von Vokabularen und Klassifikationen, Aufbau von Normdatenrepositorien oder die Verwendung von Georeferenzen - haben gezeigt, dass diese Ansätze jeweils allein kaum befriedigende Ergebnisse liefern. Vielmehr kann der komplexen Gesamtheit des Bildes wohl nur die Verbindung mehrerer Methoden gerecht

werden. Zugleich wird auch deutlich, dass weiterhin die Verbindung der Bilddateien mit (nach wie vor in Textform codierten) Inhaltskonzepten eine zentrale Aufgabe bleiben wird.

Der Beitrag wird sich auf die Frage konzentrieren, wie diese Inhaltskonzepte in möglichst pragmatischer Weise bereitgestellt werden können. Hier stellen sich Fragen der Standardisierung beziehungsweise des Einsatzes von sogenannten Normdaten.

Im Bereich der sogenannten Normdaten gibt es für Kunstwerke – im Unterschied etwa zu Personen – kaum ein flächendeckendes Angebot. Es ist auch kaum anzunehmen, dass die hierfür zuständigen Institutionen – in Deutschland etwa die DNB – dem Bedarf werden ausreichend nachkommen können. Artefakte sind gegenüber Personennormdaten aufgrund ihrer Vielgestaltigkeit grundsätzlich schwerer zu handhaben und, je nach Definition, was alles als verzeichniswürdiges Kunstwerk zu verstehen ist, unter Umständen weitaus umfangreicher in der Anzahl. Auch dort wo sich einschlägige Institutionen der Aufgabe angenommen haben, bleibt das entweder auf die nationale Dimenion beschränkt (etwa mit den Datenbanken Merimee oder Joconde in Frankreich, oder dem RKD in den Niederlanden), oder droht unweigertlich unausgewogen und fragmentarisch zu bleiben (CONA – The Cultural Objects Name Authority des Getty Research Institute). Das Deutsche Dokumentatonszentrum für Kunstgeschichte, Foto Marburg, hat zwar vielfach die Bedeutung von Werknormdaten unterstrichen, jedoch bislang keinen Vorschlag für deren Bereitstellung gemacht. Nach heutigem Ermessen kann wohl auch nicht davon ausgegangen werden, dass es möglich oder sinnvoll ist, ein vollständiges Referenzrepositorium aller Bau- und Kunstwerke anzustreben.

Wenn man von dieser Annahme ausgeht, dann ist es aber geradezu notwendig, dass jederzeit kurzfristig Datensätze für jeweils benötigte Kunstwerke erzeugt werden können. Hierfür bietet sich ein Datenrepositorium an, das von der Wikipedia-Community seit 2012 parallel zur Wikipedia aufgebaut wird. Dabei ist für die folgenden Überlegungen grundlegend, dass Wikipedia-Artikel zwar in der Regel einem Wikidata-Datensatz zugeordnet sind, dass Wikidata-Datensätze jedoch auch ohne Wikipedia-Artikel existieren können und somit auch nicht den von der Wikipedia geforderten Relevanzkriterien entsprechen müssen. Auch die – im Forschungskontext oft problematischen .- Aspekte der inhaltlichen Aktualität und Gültigkeit der Wikipedia-Artikel spielt keine Roll. Wikidata beschränkt sich auf die Speicherung von atomaren Statements, die in beliebiger Zahl, in beliebiger Reihenfolge und mit der Möglichkeit der Neudefinition von Aussageparametern im Prinzip von jeder Person erstellt werden können. Dabei steht – ebenso wie in unserem Anwendungsszenario – bei Wikidata der Gedanke der Identifizierung im Vordergrund, indem möglichst viele bereits bestehende „Identifiers“ anderer Referenzrepositorien eingegeben werden. Die Tatsache der Vielzahl solcher Repositorien (die im Bibliotheksbereich mit der VIAF-Initiative zusammengefasst werden) relativiert den im Deutschen üblichen Begriff der Normdaten ebenso wie den im Englischen üblichen des Authority File. Beide Begriffe gehen von der normativen Rolle einer Nationalbibliothek bei der Ansetzung von Personennamen und Schlagwortsystematiken aus. Im Fall von Kunstwerken ist ein normativer Ansatz, der über die bloße Bezeichnung hinausgeht (und etwa Zuschreibung, Datierung, Stilzugeörigkeit etc. festlegen wollte), eher schädlich als nützlich. Vielmehr geht es um die Identifizierung der Werke und deren Verfügbarmachung für weitere Bildrepositorien und dergleichen. Diese Funktion erfüllt Wikidata, wobei die zusätzlichen inhaltlichen Statements je nach Umständen durchaus verwendet werden können.

Wikidata kann also als eine Art Meta-Referenzrepositorium fungieren, das zudem Skalierbarkeit im kollektiven Zugriff, Internationalität und Vielsprachigkeit, sowie nicht zuletzt Nachhaltigkeit durch eine große Community bietet. Zu berücksichtigen sind freilich auch die offenen Fragen, etwa danach, welche Probleme in der Benutzbarkeit der Daten sich aufgrund der unsystematischen Struktur und der grundsätzlich nicht festgelegten Entwicklungsoptionen ergeben können.

Als Beispiel wird hierzu ein laufendes Digitalisierungsprojekt für die genannte historische Glasdiasammlung an der Humboldt-Universität als Poster vorgestellt, an dem die genannten Aspekte dargestellt werden können.

Bibliographie

Kohle, Hubertus (2013): *Digitale Bildwissenschaft*. Glückstadt.

Krause, Celia / **Reiche, Ruth** (2015): *Ein Bild sagt mehr als tausend Pixel? Digitale Forschungsansätze in den Bild- und Objektwissenschaften*. Glückstadt.

Patton, Glenn E. (2010): *Funktionale Anforderungen an Normdaten: Ein konzeptionelles Modell* (IFLA Working Group on Functional Requirements and Numbering of Authority Records - FRANAR). München.

Woitas, Kathi (2013): *Bibliografische Daten, Normdaten und Metadaten im Semantic Web – Konzepte der Bibliografischen Kontrolle im Wandel*. Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft 338. Berlin urn:nbn:de:kobv:11-100209272.

Webbasierte Morphemannotation Diachroner Korpora: Ein Weg zu mehr Nachhaltigkeit?

Peukert, Hagen
hagen.peukert@uni-hamburg.de
Universität Hamburg, Deutschland

Die Annreicherung historischer Texte mit derivationsmorphologischen Informationen ist aus der Sichtweise automatisierter Verfahren eine doppelte Herausforderung. Im Gegensatz dazu zeigt die automatische Erkennung von Flexionen bereits gute Ergebnisse (Dipper 2011, Bollmann et al 2014a,b). Die Herausforderungen lassen sich auf zwei wesentliche Unterschiede zurückführen. Erstens ist die Identifikation eines Derivationsmorphems aufgrund der vielzähligen Wortbildungsmechanismen und daraus folgender Überlappungsprobleme bei nicht agglutinierenden Sprachen algorithmisch nicht exakt zu bestimmen (vgl. Givón 1971, Dryer et al 2011, Lehmann 1973) und derzeit nur durch Abgleich mit einem a priori vorhandenen Lexikon überhaupt in Annäherung möglich. Zweitens ändert sich sowohl Form als auch Bedeutung eines Morphems über die Zeit hinweg, sodass sich daraus eine weitere Art der Überschneidung von Form sowie Inhalt (Bedeutung) einzelner Morpheme ergeben kann (vgl. Berg 1998, Faiß 1992, Kastovsky 2009). Vorausgesetzt man lässt eine Komplexitätsreduktion durch die Einführung von Zeitintervallen zu und vernachlässigt so die relativ langen Zeiträume, in denen sich Morpheme in einer Übergangsphase hin zu neuer Form und Inhalt befinden, folgt daraus immer noch, dass entsprechende Lexika für jede festgelegte Zeitperiode vorhanden sein müssen, um größere Textkorpora automatisch bearbeiten zu können. Je feinerkörniger die Zeitintervalle gewählt werden, desto größer wird die Anzahl an benötigten Lexika (proportional zur Anzahl der Zeitperioden). Feine Unterteilungen in den Zeitintervallen sind oft notwendig, um in der Folge die beobachteten Sprachwandelmechanismen genauer und ursächlich erklären zu können.

Die Lösung dieses Problems liegt demnach in der effizienten Erstellung einer entsprechenden Ressource, welche neben dem Lemma mit der Ausweisung der morphologischen Bestandteile auch die Zeit erfasst. Neben den morphologischen Informationen (Wurzel, Position und Anzahl von Präfixen und Suffixen) werden auch die Wortklasse und das Korpus erfasst. Bei Composita gehören zudem Kopf und semantische Kategorien (dvandva, bahuvrihi, appositional) zum Annotationsschema. Effizienzgewinne können dabei einerseits durch eine möglichst geschickte Aufteilung von standardisierbaren Routineaufgaben, welche Automaten abarbeiten können, und komplexeren Entscheidungsaufgaben, die ein Bearbeiter manuell treffen muss, erzielt werden. Andererseits kann dem Bearbeiter bei der Entscheidungsfindung mit der Bereitstellung von wichtigen Informationen und Komfortabilität bei der Bedienung und Präsentation geholfen werden.

Ein *Use Case* eines solchen Wortanalysewerkzeugs konnte mit dem Morphilo-Toolset als *Stand- A lone*-Anwendung ausprogrammiert werden. Diese Software berücksichtigt beim Abgleich großer Textkorpora mit dem Lexikon die Zeitspanne. Sind für das angegebene Zeitintervall Einträge vorhanden, werden diese automatisch zugewiesen. Die übrigen (unbekannten) Typen des Textes werden als neue Lemmata angelegt. Falls ein Lemma in der Vorgängerperiode bereits existiert, wird der aktuelle Eintrag mit den Informationen der Vorgängerperiode belegt und zur Bearbeitung präsentiert. Andernfalls (d.h. der Eintrag ist auch in keiner Vorgängerperiode registriert) wird das Token mit einer generischen Zerlegung automatisiert in seine morphemischen Bestandteile aufgeteilt. Der Nutzer bestätigt eine dieser Zerlegungen oder nimmt über entsprechende Menüs Änderungen vor. Erst jetzt werden diese Informationen persistent abgelegt (vgl. Peukert 2012).

Im so etablierten Workflow hat sich zunächst gezeigt, dass die Bearbeitung von englischen Texten aus dem 17. Jahrhundert (PPCMBE,

Kroch et al 2010) schnell und effizient zu bewerkstelligen ist, wenn eine kritische Masse an Einträgen bereits vorhanden ist, da das TTR mit zunehmender Textgröße gegen Null strebt, d.h. immer nur wenige unbekannte Wörter in jedem neuen Text vorzufinden sind (Baayen 1996). Dieser Effekt trat bei der Bearbeitung von frühen mittelenglischen Texten aus dem 12. Jahrhundert (PPCME2, Kroch and Taylor 2000), nicht auf. Es zeigte sich, dass die fehlenden Schreibstandards von historischen Texten die notwendige Lemmatisierung scheitern ließen und somit auch ein Abgleich mit dem Lexikon nicht gelingen konnte (vgl. Peukert 2014).

Berücksichtigt man diese beiden Erfahrungen – schnelle Annotation bei kritscher Masse an Einträgen und langsame Annoation bei fehlenden Standards – bei der Entwicklung von Lösungsstrategien, trifft man unweigerlich auf den Nachhaltigkeitsgedanken beim Ressourcenaufbau, der vorgibt, dass die kostenintensiven Annotationsaufgaben möglichst nicht mehrfach erledigt, aber nachgenutzt werden sollen. Dies impliziert eine gemeinschaftlich-synergetische Bearbeitung der Annotationszuweisung, da man die spätere Nutzung der Resource mit eigener (sehr geringer) Annotationsarbeit "bezahlen" kann. Auf diese Weise können annotierte Daten unterschiedlicher Zeiträume gesammelt werden. In der Fortführung dieser Idee ist die Architektur einer webbasierten Komponente entstanden (Abb. 1), bei der ein *Multi-User*-Design die Annotationsarbeit an unterschiedlichen Korpora verteilt und Zuweisungen aus verschiedenen Lexika aber der passenden Zeitperiode und Sprache erlaubt. Um dem Problem der fraglichen Qualität der Annotationen entgegenzuwirken, ist es möglich, die Lexika, die man zur Bearbeitung benötigt, auszuwählen. Möchte man sein eigenes Korpus mit derivationsmorphologischen Informationen anreichern lassen, fliesst die jeweils eigens geleistete Annotationsarbeit in den Gesamtdatenbestand ein. Inwiefern die gesammelten Annotationsdaten mit weiteren Verfahren hinsichtlich ihrer Qualität getestet, bewertet und weiter bearbeitet werden können, wird Gegenstand einer weitergehenden Diskussion sein.

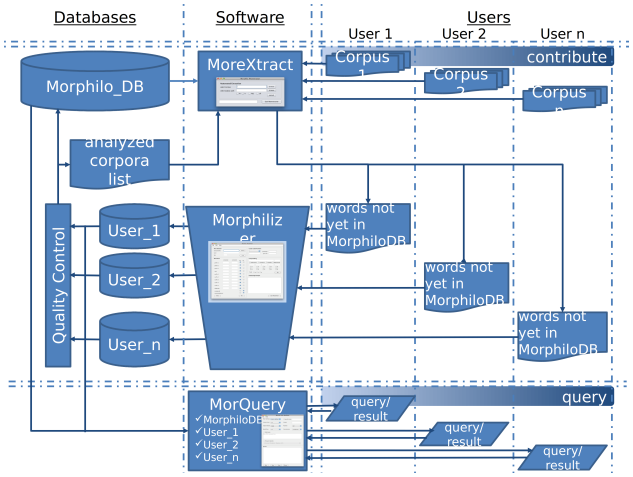


Abb. 1: Architekturentwurf zur Integration des Annotationswerkzeugs in eine webbasierte Anwendung im Mehrnutzerbetrieb

Für die Lösung der noch nicht endgültig fertiggestellten Komponente (in Abb. 1 mit "Quality Control" bezeichnet) werden statistische Verfahren in Anlehnung an das maschinelle Lernen vorgestellt, die sich an zwei unterschiedlichen Strategien ausrichten. Erstens steht die häufigkeitsbedingte Analyse gleicher oder ähnlicher Einträge der verschiedenen Datenbestände der Nutzer (User_1, ..., User_n) im Vordergrund. Diese Daten werden genutzt, um Ausreißer und falsche Annotationen mittels automatisierter statischer Signifikanztests zu identifizieren. Dieser Ansatz wird mit einer nutzerorientierten Strategie kontrastiert. Diese zweite Strategie bezieht die Verhaltensdaten der Nutzer ein, d.h. wie oft werden welche Datenbestände anderer Nutzer für die anstehende Annotation ausgewählt. Auch hier basiert der Ausschluss von vermeindlich fehlerhaften Daten mittels eines vorher festgelegten Signifikanzniveaus. Die mit einer dieser Strategie bereinigten Datenbestände könnten danach in den Hauptdatenbestand (Morphilo_DB in Abb. 1) überführt werden.

Bibliographie

Baayen, Harald (1996): „The effects of lexical specialization on the growth curve of the vocabulary", in: *Computational Linguistics* 22: 455–480.

Berg, Thomas (2009): *Structure in language: A dynamic perspective*. New York: Routledge.

Bollmann, Marcel / Petran, Florian / Dipper, Stefanie / Krasselt, Julia (2014a): „CorA: A web-based annotation tool for historical and other non-standard language data", in:

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) 86–90.

Bollmann, Marcel / Petran, Florian / Dipper, Stefanie (2014b): „Applying Rule-Based Normalization to Different Types of Historical Texts — An Evaluation“, in: Zygmunt Vetulani and Joseph Mariani (eds.): *Human Language Technology Challenges for Computer Science and Linguistics. 5th Language and Technology Conference*, LTC 2011. Revised Selected Papers. Lecture Notes in Computer Science 8387. Springer 166–177.

Dipper, Stefanie (2011): „Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison“, in: *Journal for Language Technology and Computational Linguistics*. Special Issue 26 (2): 25–37.

Dryer, Matthew S. / Haspelmath, Martin (eds.). (2011): *The world atlas of language structures Online*. München: Max Planck Digital Library.

Faiß, Klaus (1992): *English historical morphology and word-formation: Loss versus enrichment*. Trier: Wissenschaftlicher Verlag.

Givón, Talmy (1971): „Historical syntax and synchronic morphology: An archaeologist's field trip“, in: *Chicago Linguistic Society* 7: 394–415.

Kastovsky, Dieter (2009): „Diachronic perspectives“, in: Lieber, Rochelle / Štekauer, Pavol (eds.): *The Oxford handbook of compounding*. Oxford: Oxford University Press 321–340.

Kroch, Anthony / Santorini, Beatrice / Diertani, Ariel (2010): *The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE)*. Department of Linguistics, University of Pennsylvania: CD-ROM, first edition http://www.ling.upenn.edu/hist-corpora/ .

Kroch, Anthony / Taylor, Ann (2000): *The Penn-Helsinki Parsed Corpus of Middle English (PPCME)*. Department of Linguistics, University of Pennsylvania: CD-ROM, first edition http://www.ling.upenn.edu/hist-corpora/ .

Lehmann, Winfred P. (1973): „A structural principle of language and its implications“, in: *Language* 49 (1): 47–66.

Peukert, Hagen (2014): „The Morphilo Toolset: Handling the Diversity of English Historical Texts“, in: Ammermann, Anne / Brock, Alexander / Pflaeging, Jana / Schildhauer, Peter (eds.): *Facets of Linguistics: Proceedings of the 14th Norddeutsches Linguistisches Kolloquium 2013*. Frankfurt: Peter Lang 161–172.

Peukert, Hagen (2012): „From Semi-Automatic to Automatic Affix Extraction in Middle English Corpora: Building a Sustainable Database for Analyzing Derivational Morphology over Time“, in: Jancsary, Jeremy (ed.): *Empirical Methods in Natural Language Processing, Wien, Scientific series of the ÖGAI* 413–23.

Where the words are: a visual interactive exploration of plants names

Therón, Roberto
theron@usal.es
Universidad de Salamanca, Spanien

Dorn, Amelie
amelie.dorn@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Seltmann, Melanie
melanie.seltmann@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Benito, Alejandro
abenito@usal.es
Universidad de Salamanca, Spanien

Wandl-Vogt, Eveline
eveline.wandl-vogt@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Gabriel Losada Gómez, Antonio
alosada@usal.es
Universidad de Salamanca, Spanien

Wo die Wörter sind: eine visuell-interaktive Erforschung von Pflanzennamen

In den Digital Humanities werden häufig Visualisierungsmethoden eingesetzt, um bestimmte Trends, Beziehungen oder Inhalte innerhalb oder zwischen verschiedenen Datensätzen hervorzuheben. Oft werden gut etablierte und weit verbreitete Arten graphischer Darstellung von Daten herangezogen, wie Verbert (2015) gezeigt hat. Der Einsatz innovativer Visualisierungsmethoden für die Datenerforschung und den Datenzugriff ist jedoch bei Humanities-Projekten, die sich mit

nicht-numerischen Daten beschäftigen, noch relativ selten. In diesem Beitrag stellen wir ein Visualisierungstool vor, das im Rahmen des DH-Projekts *exploreAT! – exploring Austria's culture through the language glass* entwickelt wird, und erläutern dessen Anwendung am Beispiel der Pflanzennamen-Sammlung für das Wörterbuch der bairischen Dialekte in Österreich.

exploreAT! (vgl. Wandl-Vogt et al, 2015) bietet unterschiedliche Einblicke in die vielfältige Beschaffenheit der deutschen Sprache in Österreich, durch exploratives Erforschen mittels einer Synthese von digitalen Infrastrukturen, Lexikographie, visueller Analyse und Citizen Science. Das Projekt basiert auf einer Sammlung von Daten zu den bairischen Dialekten in Österreich aus dem frühen 20. Jahrhundert aus der Region der ehemaligen österreichisch-ungarischen Monarchie. Die Datenerhebung erfolgte ursprünglich mittels Fragebögen, die eine Vielzahl von Themen aus dem Alltag abdecken. Die gesammelten Daten bestehen aus rund 200.000 Stichwörtern in geschätzten 4 Millionen Datensätzen. Teile davon wurden als fünfbändiges Wörterbuch mit etwa 50.000 Stichwörtern (WBÖ), und Teile als Datenbank (DBÖ) ausgegeben. Innerhalb des Projekts gibt es vier spezifische, aber miteinander verbundene Arbeitsbereiche: kulturelle Lexikographie, semantisch-technologieorientierte Forschungsinfrastrukturen, visuelle Analyse und Bürgerwissenschaften. Des Weiteren werden use-cases für spezifische Themen wie Pflanzennamen, Farben oder Lebensmitteln entwickelt. TEI / XML Schnittstellen werden eingesetzt, um die Organisation von Metadaten, Konzepten und linguistischen Daten zu verbessern. Darüber hinaus ist vorgesehen, weitere Zugangspunkte zur Arbeit mit LOD zu schaffen, ontologische Ressourcen zu nutzen und damit die Visualisierung von konzeptionellen und semantischen Informationen zu gewährleisten.

Mit Hilfe des vorgestellten visuellen Analysetools werden weitere Einblicke in die komplexe Struktur dieser Dialektdaten gegeben, wobei ein intuitiver und leicht zugänglicher Ansatz vorgesehen ist. In diesem Beitrag nehmen wir Pflanzennamen als exemplarischen Fall für die visuelle Exploration, Analyse und Darstellung von Datenstrukturen.

Der Prototyp dieses Tools basiert auf einer Treemap-Visualisierungsmethode (vgl. Shneiderman, 1992), da diese eine kompakte Art und Weise für die Übertragung von Hierarchien ermöglicht. Der Zweck des Tools besteht darin, ein Mittel zur interaktiven Erforschung der verfügbaren Daten bereitzustellen, so dass der Benutzer Verständnis dafür gewinnt, wie das Wissen, das sich auf ein bestimmtes Wort (oder eine Zeichenkette) bezieht, in der Datenbank "gespeichert" ist, wobei die jeweiligen Lemmata mit der Benutzerabfrage zusammenhängen. Abgesehen von der Darstellung des resultierenden Sets von Lemmata, bauen wir eine Hierarchie je nach Kontext der Lemmata (in diesem Fall sind wir daran interessiert, die Lemmata in Bezug auf verschiedene Pflanzenarten zu gruppieren). Deshalb verwenden wir die beiden wichtigsten visuellen Merkmale von Treemaps: a) das Treemap-Layout (basierend auf einem Satz von verschachtelten Rechtecken, wobei jedes Rechteck einen Zweig der Hierarchie darstellt, der dann mit kleineren Rechtecken, welche Unterzweige darstellen, gekachelt wird) und b) die Fläche jedes Rechtecks (die proportional zur Größe der Daten ist).

Da in diesem Prototyp Pflanzennamen von größter Bedeutung sind, aber dem Benutzer, der eventuell mit den wissenschaftlichen Namen der Pflanzen nicht vertraut ist, wichtige Information verborgen bleiben könnte, entschieden wir uns für eine visuelle Art den Kontext (Pflanzen) der Lemmata, die in Zusammenhang mit der Abfrage stehen, zu vermitteln: wir verwenden den Flickr-Webdienst, um Fotos abzurufen, die mit dem wissenschaftlichen Namen der Pflanze versehen sind (siehe Abbildung 1).

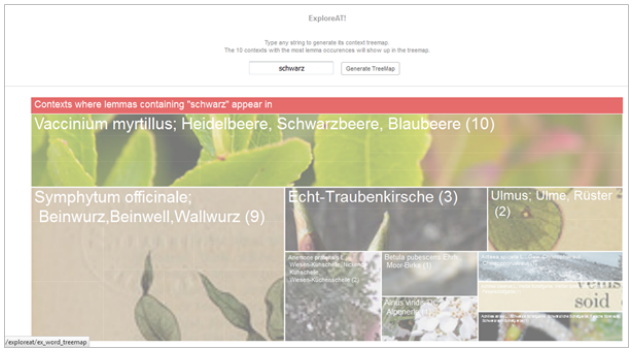


Abbildung 1: Beispiel für eine visuelle Darstellung von Pflanzennamen mit verschachtelten Rechtecken.

Als Ergebnis unseres visuellen Ansatzes kann der Benutzer die Verteilung der Lemmata in Abhängigkeit von den Pflanzen, auf die sie sich bezieht, verstehen (jedes Rechteck enthält das abgerufene Foto einer bestimmten Pflanze mit einem Bereich entsprechender Größe, die davon abhängt, wie viele Lemmata in Zusammenhang damit stehen). Der Benutzer kann dann auf das Rechteck seiner Wahl klicken, um tiefer zu gehen

und alle relevanten Informationen für die mit dieser Pflanze zusammenhängenden Lemmata zu erhalten (siehe Abbildung 2).

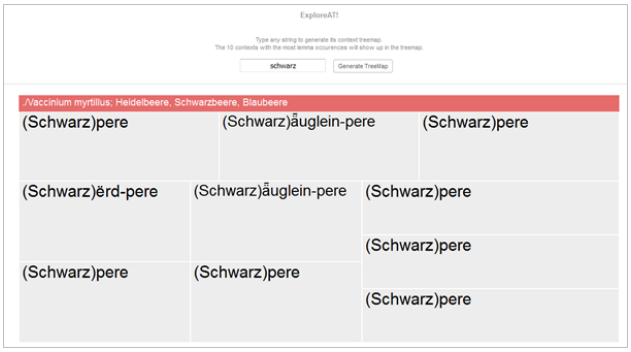


Abbildung 2: Beispiel für die Exploration von Pflanzennamen-Lemmata in einem bestimmten Rechteck, in diesem Fall Vaccinium myrtillus; Heidelbeere, Schwarzbeere, Blaubeere (siehe Abbildung 1).

Schließlich öffnen sich künftige Arbeitsfelder dank der Tatsache, dass dieser visuelle Ansatz auch noch gültig ist, wenn wir die Lemmata nach anderen Kriterien (d.h. nach einer mehrstufigen Hierarchie) gruppieren. Zum Beispiel könnte man zuerst die Lemmata nach Pflanze gruppieren; dann könnte man für eine bestimmte Pflanze die dazugehörigen Lemmata nach Zeit gruppieren, die wiederum nach Regionen gruppiert werden. Mit diesen verschiedenen Arten der Gruppierung können diverse andere Daten mit einer ähnlichen strukturellen Beschaffenheit in derselben Weise visualisiert und analysiert werden. Dies würde unser Tool vielseitig und auch offen für andere Daten, nicht nur Pflanzennamen, machen.

Bibliographie

Verbert, Karen (2015): „On the Use of Visualization for the Digital Humanities“ in: *DH2015: Global Digital Humanities*.

Wandl-Vogt, Eveline / Kieslinger, Barbara / O'Connor, Alexander / Theron, Roberto (2015): „exploreAT! Perspektiven einer Transformation am Beispiel eines lexikographischen Jahrhundertprojekts“, in: *DHd 2015: Von Daten zu Erkenntnissen*.

Shneiderman, Ben (1992): „Tree visualization with tree-maps: 2-d space-filling approach“, in: *ACM Transaction on Graphics (TOG)* 11 (1): 92-99.

Zukünftiger Teil eines Fachinformationsdienstes: Eine Datenbank zur *Fachgeschichte der deutschsprachigen Musikwissenschaft zwischen ca. 1810 und ca. 1990*, projektiert am Max-Planck-Institut für empirische Ästhetik, Frankfurt am Main

van Dyck-Hemming, Annette
annette.van-dyck-hemming@aesthetics.mpg.de
Max Planck-Institut für empirische Ästhetik, Deutschland

Die Musikwissenschaft unter dem Namen ‚Musikwissenschaft' ist eine relativ junge Disziplin an den Universitäten: erst ab den zwanziger Jahren des 19. Jh. wurden ihr Lehrstühle zugestanden und nicht vor den 1880er Jahren Institute gegründet. Sie erlebte einen großen Aufschwung nach der Jahrhundertwende und ließ sich in weiten Teilen instrumentalisieren im Nationalsozialismus. Wie andere Disziplinen hat auch die Musikwissenschaft lange gebraucht, um sich die Selbstverständlichkeit fachgeschichtlicher Selbstreflexion zuzugestehen (Gerhard 2000). Seit den 1990er Jahren jedoch wachsen Forschungsinteresse und -output sehr deutlich an.

Davon inspiriert wurde 2014 am neu gegründeten Max-Planck-Institut für empirische Ästhetik in Frankfurt am Main ein Projekt konzipiert und etabliert, das zum Einen die Pflege eines aktuellen Forschungsnetzwerkes zum Thema ‚Fachgeschichte in der Musikwissenschaft', zweitens die Förderung von Einzelstudien und drittens die Bereitstellung von Quellen und Forschungsergebnissen vorsieht. Als vierter Teil des Projektes ist geplant, mit der in der Musikwissenschaft relativ neu erarbeiteten Basis fachhistorischer Daten und vor dem Hintergrund der neueren und älteren

soziologischen Netzwerkforschung die Daten von fachgeschichtlich in Erscheinung getretenen Personen und Institutionen (Universitäten, Akademien, Vereine, Verlage etc.) zu sammeln und miteinander sowie zu Art und Menge der mit diesen Daten zusammenhängenden Veröffentlichungen in Beziehung setzen (van Dyck-Hemming/ Wald-Fuhrmann 2016).

Das Projekt dient der unumkehrbaren Verankerung historiographischer Reflexion in der Musikwissenschaft und so der Erweiterung und Verwissenschaftlichung musikologischer Zugänge. Im Sinne von Ludwik Fleck (1935) und Thomas Kuhn (1967), die vom notwendigen Zusammenhang zwischen dem Inhalt einer Wissenschaft und ihren historischen Erkenntnisprozessen ausgingen, soll die musikhistoriographische Datenbank kein Leistungsindex und keine Ahnentafel der Musikwissenschaft werden, sondern eine valide und nachprüfbare Datenbasis zusammenstellen, die präzise Darstellungen von Prozessen, Netzwerken und Verteilungen ermöglicht.

Mittels der sich Standards und Normdaten zunutze machenden, relationalen Datenbank werden Thesen generiert werden können unter anderem in Bezug auf Fragen nach der Existenz und Art von Personennetzwerken in der Musikwissenschaft, nach Kontinuitäten oder Veränderungen musikwissenschaftlicher Forschungspräferenzen, nach Abhängigkeiten zwischen zeitgeschichtlichen Veränderungen und der Institutionalisierung einer neuartigen Wissensdisziplin. Die Ergebnisse sollen anschaulich visualisiert und dazu auch in Zeit und (historischen) Raum dimensioniert werden. Besonderer Wert wird gelegt auf technisch niedrigschwellige Benutzeroberflächen der schließlich öffentlich zugänglichen Datenbank bei gleichzeitig hohem Anspruch an Transparenz und Überprüfbarkeit von Quellen und Verfahren. Teil des Konzepts ist auch, dass die weitere Befüllung durch Fachwissenschaftlerinnen und Fachwissenschaftler unter redaktioneller Moderation erfolgen kann.

Als natürliche Partner dieses Projektes haben sich Bibliotheken erwiesen: Durch mit der institutseigenen Bibliothek und den Verfahren der Deutschen Nationalbibliothek abgestimmte Workflows wird sichergestellt, dass die mit fachgeschichtlichem Filter gewählten Personen und Institutionen Bestandteile der Gemeinsamen Normdatei sind; gegebenenfalls werden als Ergebnis unserer Recherchen GND-Datensätze korrigiert, ergänzt oder erstellt.

Über die GND hinausgehende Informationen wie Relationen und Beziehungsbeschreibungen etc. nimmt unsere Datenbank außerdem auf. Alle Datensätze werden auf die Quellen rückführbar sein; bislang in Form bibliographischer Nachweise.

Forschungspräferenzen sollen hauptsächlich über die Auswertung der Schlagworte und Klassifikationen von musikwissenschaftlichen Publikationen erfasst werden. Diese Bibliotheksdaten stellt uns die Musikabteilung der Bayerischen Staatsbibliothek (BSB) zur Verfügung. Mit ihr werden das Datenmodell sowie technische Voraussetzungen und Entscheidungen abgestimmt und hinsichtlich der Realisierung auf allen Ebenen kooperiert: Die BSB führte seit Jahrzehnten den Sammelschwerpunkt Musikwissenschaft, besitzt umfassende und intersubjektiv abgesicherte Kompetenz in der Formal- und Inhaltserschließung von Publikationsdatensätzen und hat in diesem Rahmen auch bereits Vorarbeiten zu einer Ontologie der Musikwissenschaft geleistet, an die wir uns anschließen wollen. Für die Fachöffentlichkeit stellt die BSB seit einigen Jahren das Webportal und die Infrastruktur ‚Fachinformationsdienst Musik' (https://www.vifamusik.de) zur Verfügung. In diesem Rahmen soll auch die Datenbank zur Fachgeschichte der Musikwissenschaft implementiert und insgesamt oder in Teilen von der BSB gehostet werden. Ähnlich dem Suchportal BSB opac plus – eine Eigenentwicklung der BSB (https://opacplus.bsb-muenchen.de/) – könnte die fachgeschichtliche Datenbank funktionieren – mit erweiterten Funktionen und sinnvollen Visualisierungsmöglichkeiten. Und wie im Fall eines Bibliothekssuchportals wird die Perspektive auf weitere Anwendungen außerhalb der Musikwissenschaft berücksichtigt.

Das Projekt am MPIEA ist mit einer WissMA-Stelle für 10 Jahre sowie Hilfskraft-Stellen ausgestattet; geleitet wird es aber von der unbefristet eingesetzten Direktorin der Abteilung Musik. Nachhaltige öffentliche Verfügbarkeit des in 10 Jahren zu erarbeitenden Datenbestandes verspricht darüber hinaus das Einpflegen von Teilen der DB in die Gemeinsame Normdatei, die präventive Kooperation mit der Bayerischen Staatsbibliothek als mindestens ebenso langfristig ausgerichteter und in nachhaltiger Datenpflege erfahrene Institution und die Implementierung des Webzugangs im Rahmen eines Fachinformationsdienstes. Über die Fertigstellung hinausreichende Datenaktualität und beständige Erweiterung des Datenbestandes erwarten wir uns von der

Bereitstellung eines Zugangs für registrierte Musikwissenschaft Treibende.

Projektleitung: Dr. Melanie Wald-Fuhrmann (Direktorin Abteilung Musik), Projektkoordination: Dr. Annette van Dyck-Hemming (wiss. MA)

Bibliographie

Gerhard, Anselm (ed.) (2000): *Musikwissenschaft – eine verspätete Disziplin*. Stuttgart.

van Dyck-Hemming, Annette / Wald-Fuhrmann Melanie (2016): „Vom Datum zum historischen Zusammenhang. Möglichkeiten und Grenzen einer fachgeschichtlichen Datenbank“, in: Bolz, Sebastian / Kelber, Moritz / Knoth, Ina / Langenbruch, Anna (eds.): *Wissenskulturen der Musikwissenschaft. Generationen – Netzwerke – Denkstrukturen*. Bielefeld: transcript 261-278.

Fleck, Ludwik (2015): *Entstehung und Entwicklung einer wissenschaftlichen Tatsache: Einführung in die Lehre vom Denkstil und Denkkollektiv*. Frankfurt am Main 10. Auflage [Basel 1935].

Kuhn, Thomas S. (2014): *Die Struktur wissenschaftlicher Revolutionen*. Frankfurt am Main 24. Auflage.

Zwei grundlegende Fragen der digitalen Nachhaltigkeit: Wie können wir die heterogenen Forschungsfragen und die Community bei der Verfügbarmachung von Forschungsdaten miteinbeziehen?

Odebrecht, Carolin
carolin.odebrecht@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Dreyer, Malte
malte.dreyer@cms.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Lüdeling, Anke
anke.luedeling@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Krause, Thomas
krauseto@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Zielstellung

Digitale Nachhaltigkeit verstehen wir als eine komplexe Anforderung, die aus besonders vielen Blickwinkeln betrachtet werden kann und sollte. Wir stellen mit dem LAUDATIO-Projekt [1] eine Möglichkeit vor, digitale Nachhaltigkeit herzustellen, in dem wir einen unabhängigen und freien Zugriff auf historische Korpora zum Zweck der Wiederverwendung für Forschungszwecke, die über diejenigen hinausgehen, für die die Daten ursprünglich gesammelt wurden, ermöglichen. Eine so definierte digitale Nachhaltigkeit realisieren wir, in dem wir eine Vielzahl an verschiedenen, teils sehr unterschiedlichen historischen Korpusdaten inklusive einer umfangreichen aber einheitlichen Dokumentation, die ihre Erschließbarkeit in Bezug auf konkrete Nutzungsszenarien gewährleistet, bereitstellen. Aus dieser Arbeit zeigt sich, dass eine interdisziplinäre und insbesondere lokale institutionelle Zusammenarbeit mit den Korpuserstellern notwendig ist, die einen engen kommunikativen Austausch von Zielen und Anforderungen diesbezüglich sowie eine Identifikation von möglichen Kooperationen erst ermöglicht. Digitale Nachhaltigkeit definiert sich dementsprechend aus der jeweiligen Perspektive der Daten, der Dokumentation und der Institutionen ein wenig anders. Diese drei Perspektiven und deren Zusammenspiel in Bezug auf die digitale Nachhaltigkeit wollen wir anhand von historischen Korpora (vgl. Claridge 2008; Kytö 2011; Gippert und Gehrke 2015) diskutieren und an einem Best-Practice-Beispiel einer interdisziplinären Zusammenarbeit der Philosophischen Fakultät und dem Rechenzentrum der Humboldt-Universität zu Berlin (HU) erklären.

Heterogenität von Forschungsdaten

Die vielfältige Datenlandschaft in den Geisteswissenschaften stellt eine große Herausforderung in Bezug auf die digitale Nachhaltigkeit dar, da es unterschiedliche Datenmodelle und -formate und verschiedene Aufbereitungs- und Analyseverfahren gibt, die alle aus immer neuen Nutzungsszenarien und ihren Forschungsfragen, resultieren. In diesem Sinne verstehen wir die Arbeit mit Korpora als eine innovative, fortlaufende, wissenschaftliche Arbeit, die sich nicht ausschließlich auf existierende Standards stützen kann (auch wenn solche Standards natürlich immer beachtet werden müssen). Es werden daher zusätzlich andere, neue Forschungsdatenmodelle und -formate sowie Ressourcentypen entwickelt und auf die unterschiedlichsten Weisen weiter- und wiedergenutzt. Beispielsweise werden in den Projekten, die LAUDATIO unterstützt, eine Vielzahl an allein XML-basierten Formaten (nach z.B. Dipper 2005; Schmidt und Wörner 2009; Romary et al. 2015; TEI Consortium 2015) sowie CSV-basierte Formate (nach z.B. Nivre, Hall und Nilsson 2004; Krause und Zeldes 2016) für die Erstellung von historischen Korpora genutzt. Daneben finden auch und graphbasierten Lösungen (nach z.B. Ide und Suderman 2014) sowie proprietäre Formate, wie bei Ágel und Hennig (2007) [2], und immer häufiger JSON-basierte Formate wie bei Vertan et al. (2016) Anwendung. Zusätzlich dazu können Korpora in mehreren Formaten vorliegen, vor allem wenn Korpora in einer Mehrebenenarchitektur (vgl. Romary und Ide 2004; Lüdeling 2012) entworfen und verschiedene Formate für unterschiedliche Abschnitte des Forschungsdatenzyklus (vgl. Rümpel 2011) genutzt werden.

Der innovative Charakter der Korpuserstellung zeigt sich neben den verwendeten Formaten auch in den Annotationsrichtlinien: Einige Formate wie TIGER-XML (Romary et al. 2015) und tcf (Heid et al. 2010) legen die Anzahl und die Bedeutung der Annotationen inklusive ihrer Tagsets fest, andere wie TEI-XML (TEI Consortium 2015) lassen Spielraum in der Serialisierung und in der Anwendung, in dem es beispielsweise mehrere valide Möglichkeiten gibt, Autoren in einem Dokument auszuweisen. [3] Wieder andere Formate wie EXMARaLDA-XML (Schmid und Wörner 2009) oder PAULA-XML (Dipper 2005) geben keinerlei solcher Beschränkungen vor.

Insbesondere bei solchen Formaten mit größtmöglichem Spielraum für Interpretationen in Form von Annotationen zeigt sich die Vielfältigkeit der korpusbasierten Forschung. Ein typisches Annotationsbeispiel für historische Korpora sind Normalisierungen. Viele Korpora besitzen eigene Richtlinien für die Normalisierung, vgl. zum Beispiel für historisches Deutsch Jurish (2010), Bollmann et al. (2012); Odebrecht et al., (eingereicht). Auch für verschiedene Forschungsfragen wesentliche Kategorisierungen wie Wortarten gibt es annähernd für jedes historische Korpus eine eigene Lösung. So werden beispielsweise de-facto-Standards wie das STTS (Schiller et al. 1999) jeweils für ein Korpus angepasst, z.B. Dipper et al. (2013) [4] oder Fürstinnenkorresondenzkorpus [5].

Diese Beispiele zeigen, dass sich Korpora desselben Formats und ein Korpus, das in verschiedenen Formaten vorliegt, in Bezug auf ihre Annotationen stark unterscheiden können.

Ausgehend von dieser Datenlage erscheint die Nachvollziehbarkeit des Lebenszyklus der Korpora als ein weiterer wesentlicher Faktor für die digitale Nachhaltigkeit. Welche der bisher genutzten Formate und Annotationskonzepte sich langfristig durchsetzen, welche Formate wie technisch unterstützt und welche neuen Lösungen entwickelt werden, hängt dann im Wesentlichen von der Entwicklung der korpusbasierten Forschung ab. Daher setzen wir eine Archivierung und Dokumentation aller verwendeten Formate und Annotationen eines Korpus in LAUDATIO um.

Metadaten

Über einheitliche extensive Metadaten dieser heterogenen Korpusdaten und deren Lebenszyklus kann eine umfangreiche Korpusdokumentation sowie eine einheitliche Suche und ein gezielter Zugriff über eine Plattform auf die Forschungsdaten erstellerunabhängig gewährleistet werden (Bird und Simons 2001; Broeder et al. 2010; Burnard 2013; Hedges et al. 2013; Odebrecht et al. 2015). Die relevanten Kriterien für die Dokumentation und die Suche leiten sich aus den Wiederverwendungsszenarien ab (Odebrecht 2014). Um eine überfachliche Suche zu ermöglichen, wird in LAUDATIO eine technisch-abstrakte Modellierung der Metadaten eingesetzt, um die fachspezifischen Konzepte von Korpora überfachlich abzubilden (Odebrecht 2015). Neben den deskriptiven

Metadaten sind für eine nachhaltige Vorhaltung von Forschungsdaten auch administrative, strukturelle, technische und Archivierungsmetadaten relevant (vgl. Xie und Matusiak 2016; Solodovnik 2011; NISO 2004), die für eine technische Infrastruktur berücksichtigt werden müssen, um die Nachvollziehbarkeit über längere Zeiträume und wechselnde Anwendergruppen hinweg gewährleisten zu können. So ermöglicht das einheitliche Metadatenmodell eine extensive und transparente Informationsarchitektur für die unterschiedlichen Ressourcen, was wiederum ein Baustein für deren digitale Nachhaltigkeit darstellt.

Institutionelle (Zusammen-)Arbeit

Bedingt durch die unterschiedlichen Sichtweisen und Anforderungen zur digitalen Nachhaltigkeit aus den Fachdisziplinen und durch die technische Infrastruktur sind Lösungen nur im Team und in Zusammenarbeit unterschiedlicher Kompetenzen plan- und erstellbar. Im Beispiel ist für den erstellerunabhängigen Zugang zu historischen Korpora die enge Zusammenarbeit zwischen den FachwissenschaftlerInnen und LAUDATIO erforderlich. Der Computer- und Medienservice (CMS) und die Arbeitsgruppe Korpuslinguistik am Institut für deutsche Sprache und Linguistik der HU setzen mit dem LAUDATIO-Projekt einen Schwerpunkt auf eine enge institutionelle Zusammenarbeit mit den verschiedenen Arbeitsgruppen der philosophischen Fakultät, in dem es die Anforderungen für die digitale Nachhaltigkeit von Forschungsdaten der Fakultät mit den ForscherInnen erarbeitet und umsetzt und in Bezug auf die Hochschule als Ganzes in einen Entwicklungsrahmen einbettet (Dreyer und Vollmer 2016). Der Betrieb des LAUDATIO-Repositoriums wird nach Ende des Projektes, in dem Entwicklungen und Anpassungen am System vorgenommen werden, weiterhin durch das CMS sowie durch die Arbeitsgruppe Korpuslinguistik am Institut für deutsche Sprache und Linguistik gewährleistet. Um die heterogene Datenlandschaft zu verstehen, die umfassende Dokumentation zu erstellen und die neuen Entwicklungen aufzunehmen, erweist sich eine lokale Zusammenarbeit über die Fakultäten hinweg als sehr vorteilhaft. So bestehen enge Kooperationen unter anderem mit Projekten

- der Sprachgeschichte: DDD-AD [6], HIPKON [7], DDB [8], Fürstinnenkorrespondenzkorpus [9]
- der Germanistik: Märchenkorpus [10] und RIDGES [11]
- der Slawistik: “Korpuslinguistik und diachrone Syntax: Subjektkasus, Finitheit und Kongruenz in slavischen Sprachen” [12]

die jeweils sehr unterschiedliche Arbeitsweisen und Zielrichtungen haben. Durch diese Synergiebildung können Projekte ohne gesonderte Finanzierung oder Ressourcen nach ihrem Projektende durch eine genaue Kenntnis der Forschungsdatenlandschaft in einer Institution identifiziert und unterstützt werden. Die durch die Förderer (z.B. Deutsche Forschungsgemeinschaft 2015) oder Universitäten (z.B. Humboldt-Universität zu Berlin 2014) vorgegeben Richtlinien zur Veröffentlichung und Archivierung von Projektergebnissen können so ebenfalls berücksichtigt werden.

Mit diesem Ansatz können gleichzeitig zwei Ziele erreicht werden: Die Projekte müssen sich keinen umfangreichen semantischen Anforderungen unterwerfen und können sich frei ausdrücken. Gleichzeitig können sie ihre fachspezifischen Anforderungen direkt an die Arbeitsgruppe Korpuslinguistik bzw. an den CMS richten. Andersherum können die Korpusprojekte in Umfragen zur gewünschten Anforderungen und Softwarelösungen direkt befragt und in die Entwicklung mit einbezogen werden. [13] Auf diese Art findet ein Community-Aufbau statt, der sich nicht nur über gemeinsame technische Plattformen definiert, sondern rein über die inhaltliche Gemeinsamkeit der Arbeit mit historischen Korpora und somit auch über disziplinäre Grenzen hinweg.

Einordnung und Schlussfolgerungen

Der hier vorgestellte Weg, digitale Nachhaltigkeit von Forschungsdaten zu ermöglichen, stützt sich auf eine Spezialisierung auf einen bestimmten Typ von Forschungsdaten - historisches Korpus - und grenzt sich so von Ansätzen wie Zenodo [14] und dem Virtual Language Observatory (Van Uytvanck 2012) ab, die keine deutliche Eingrenzung hinsichtlich der Daten und deren Nutzungsszenarien machen.

Weiterhin zielt unsere Strategie auf die Unterstützung der Diversität der genutzten

Formate und Konzepte, die sich von den TEI spezialisierten Ansätzen wie dem Deutschen Textarchiv (Geyken 2013) und Textgrid (Hedges et al. 2013) unterscheiden.

Die Korpusersteller selbst nutzen LAUDATIO auch mehr und mehr, um ihre eigenen neuen Versionen der historischen Korpora und damit ihren wissenschaftlichen Fortschritt zu veröffentlichen. Dass unser Ansatz, sich auf die erstellerunabhängige Wiederverwendung von Korpora als eine Strategie für digitale Nachhaltigkeit zu fokussieren, auch außerhalb der Institution funktioniert, zeigt Dumont (2016).

Um die unterschiedlichen Entwicklungen und Innovationen bei der Korpuserstellung zu identifizieren, kennenzulernen und zu dokumentieren, ist eine enge, auf eine rein fachliche Ebene bezogene Zusammenarbeit mit der Community notwendig, die wir angefangen haben, im Rahmen der Philosophischen Fakultät der HU aufzubauen.

Fußnoten

1. LAUDATIO steht für **L**ong-term **A**ccess and **U**sage of **D**eeply **A**nnotated Informa **tio**n. www.laudatio-repository.org . Zugriff am 16.08.2016.
2. Ágel, Vilmos; Hennig, Mathilde; KAJUK (Version 1.1), Justus-Liebig-Universität Gießen. http://www.uni-giessen.de/kajuk/index.htm.
3. Wie es zum Beispiel mit den Element <author> möglich ist, vgl. http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-author.html Zugriff am 19.08.2016.
4. Donhauser, Karin; Gippert, Jost; Lühr, Rosemarie; ddd-ad (Version 0.1), Humboldt-Universität zu Berlin. https://referenzkorpusaltdeutsch.wordpress.com/. http://hdl.handle.net/11022/0000-0000-7FC2-7
5. Lühr, Rosemarie; Faßhauer, Vera; Prutscher, Daniela; Seidel, Henry; Fuerstinnenkorrespondenz (Version 1.1), Universität Jena, DFG. http://www.indogermanistik.uni-jena.de/Web/Projekte/Fuerstinnenkorr.htm.
6. http://www.deutschdiachrondigital.de/ Zugriff am 16.08.2016.
7. https://www.linguistik.hu-berlin.de/de/institut/professuren/sprachgeschichte/forschung/sfb632-informationsstruktur Zugriff am 16.08.2016.
8. http://korpling.german.hu-berlin.de/ddb-doku/index.htm Zugriff am 16.08.2016.
9. http://dwee.eu/Rosemarie_Luehr/?Projekte___DFG-Projekte___Fruehneuzeitliche_Fuerstinnenkorrespondenz_im_mitteldeutschen_Raum Zugriff am 16.08.2016.
10. Die Erstellung des Korpus erfolgte im Rahmen von universitärer Lehre http://www.textbewegung.de/lehre.html Zugriff am 16.08.2016.
11. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/ Zugriff am 16.08.2016.
12. https://www.slawistik.hu-berlin.de/de/member/meyerrol/subjekte/corpora Zugriff am 16.08.2016.
13. Zusätzlich helfen uns auch Evaluationen mit Kooperationspartner durch Dritte, die angebotenen Lösungen zu verbessern (z.B. Stiller et al. 2016).
14. http://zenodo.org Zugriff am 19.08.2016.

Bibliographie

Ágel, Vilmos / **Hennig, Mathilde** (eds.) (2007): *Zugänge zur Grammatik der gesprochenen Sprache*. Germanistische Linguistik 269. Tübingen: Niemeyer.

Bird, Steven / **Simons, Gary** (2001): „The OLAC Metadata Set and Controlled Vocabularies“, in: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* 7–18 arXiv:cs/0105030v1.

Bollmann, Marcel / **Dipper, Stefanie** / **Krasselt, Julia** / **Petran, Florian** (2012): „Manual and semi-automatic normalization of historical spelling - case studies from Early New High German“, in: *Proceedings of KONVENS 2012* 342–350 http://www.oegai.at/konvens2012/proceedings/51_bollmann12w/ [letzter Zugriff 22 August 2016].

Broeder, Daan / **Kemps-Snijders, Marc** / **Van Uytvanck, Dieter** / **Windhouwer, Menzo** / **Withers, Peter** / **Wittenburg, Peter** / **Zinn, Claus** (2010): „A data category registry- and component-based metadata framework“, in: *Proceedings of LREC 2010* 43–47.

Burnard, Lou (2013): „The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure“, in: *Journal of the Text Encoding Initiative* 5: 1–13 10.4000/jtei.811.

Claridge, Claudia (2008): „Historical Corpora“, in: Lüdeling, Anke / Kytö, Merja (eds): *Corpus Linguistics. An International Handbook* 1. Berlin: De Gruyter 242–259.

Dipper, Stefanie (2005): „XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation“, in: *Proceedings of Berliner XML Tage* 39–50.

Dipper, Stefanie / **Donhauser, Karin** / **Klein, Thomas** / **Linde, Sonja** / **Müller, Stefan** /

Wegera, Klaus-Peter (2013): „HiTS. Ein Tagset für historische Sprachstufen des Deutschen“, in: Zinsmeister, Heike / Heid, Ulrich / Beck, Kathrin (eds.): *Das Stuttgart-Tübingen Wortarten-Tagset: Stand und Perspektiven*. Journal for Language Technology and Computational Linguistics 28(1) 85–137.

Deutsche Forschungsgemeinschaft (2015): *Leitlinien zum Umgang mit Forschungsdaten*. Bonn: DFG. http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf [letzter Zugriff 22. August 2016].

Dreyer, Malte / Vollmer, Andreas (2016): „An Integral Approach to Support Research Data Management at the Humboldt-Universität zu Berlin“, in: *Proceedings of the European University Information Systems organisation* 320–327.

Dumont, Stefan (2016): „Fürstinnenkorrespondenzen. Experiment einer Nachnutzung“, in: *Entwicklung und Nutzung interdisziplinärer Repositorien für historische textbasierte Korpora. DHd 2016 Workshop* http://www.laudatio-repository.org/laudatio/workshop-dhd2016/ [letzter Zugriff 22. August 2016].

Gippert, Jost / Gehrke, Ralf (eds.) (2015): *Historical Corpora*. Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 5. Tübingen: Narr.

Geyken, Alexander (2013): „Wege zu einem historischen Referenzkorpus des Deutschen. das Projekt Deutsches Textarchiv“, in: Hafemann, Ingelore (ed.): *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*. Internationale Tagung des Akademienvorhabens "Altägyptisches Wörterbuch“ an der BBAW. Thesaurus Linguae Aegyptiae, 4: 221–234.

Hedges, Mark / Neuroth, Heike / Smith, Kathleen M. / Blanke, Thomas / Romary, Laurent / Küster, Marc / Illingworth, Malcom (2013): „TextGrid, TEXTvre, and DARIAH. Sustainability of Infrastructure for Textual Scholarship“, in: *Journal of the Text Encoding Initiative* 5: 1–13.

Heid, Ulrich / Schmid, Helmut / Eckart, Kerstin / Hinrichs, Erhard W. (2010): „A Corpus Representation Format for Linguistic Web Services. The D-SPIN Text Corpus Format and its Relationship with ISO Standards“, in: *Proceedings of the Seventh International Conference on Language Resources and Evaluation* 494–499.

Humboldt-Universität zu Berlin (2014): *Grundsätze zum Umgang mit Forschungsdaten an der Humboldt-Universität zu Berlin*. Unter Mitarbeit von Elena Simukovic https://www.cms.hu-berlin.de/de/ueberblick/projekte/dataman/hu-fdt-policy/view [letzter Zugriff 6. Juni 2016].

Ide, Nancy / Sudermann, Keith (2014): „The Linguistic Annotation Framework. a standard for annotation interchange and merging“, in: *Language Resources and Evaluation* 48 (3): 395–418.

Jurish, Bryan (2010): „More than Words: Using Token Context to Improve Canonicalization of Historical German“, in: *Journal for Language Technology and Computational Linguistics* 25 (1): 23–40.

Krause, Thomas / Zeldes, Amir (2016): „ANNIS3. A new architecture for generic corpus query and visualization“, in: *Digital Scholarship in the Humanities* 31 (1): 118–139 10.1093/llc/fqu057.

Kytö, Merja (2011): „Corpora and historical linguistics“, in: *Revista Brasileira de Linguistica Aplicada* 11: 417–457 10.1590/S1984-63982011000200007.

Lüdeling, Anke (2012): „A corpus-linguistics perspective on language documentation, data, and the challenge of small corpora“, in: Seifart, Frank / Haig, Geoffrey / Himmelmann, Nikolaus P. / Jung, Dagmar / Margetts Anna / Trilsbeek, Paul (eds.): *Potentials of Language Documentation. Methods, Analyses, and Utilization* 4. Language Documentation & Conservation Special Publication 3. Hawaii: University of Hawai‘i Press 32–38.

NISO (2004): *Understanding Metadatada*. Bethesda: NISO Press http://www.niso.org/publications/press/UnderstandingMetadata.pdf [letzter Zugriff 13. Februar 2015].

Nivre, Joakim / Hall, Johan / Nilsson, Jens (2004): „Memory-Based Dependency Parsing“, in: *Proceedings of the Eighth Conference on Computational Natural Language Learning* 49–56.

Odebrecht, Carolin / Belz, Malte / Zeldes, Amir / Lüdeling, Anke / Krause, Thomas (eingereicht): *RIDGES Herbology - Designing a Diachronic Multi-Layer Corpus*. Vorversion https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/download-files/pubs/odebrechtetaleingereicht_ridgesherbology.pdf/at_download/file [letzter Zugriff 22. August 2016].

Odebrecht, Carolin (2014): „Modeling Linguistic Research Data for a Repository for Historical Corpora“, in: *DH2016: Book of Abstracts* 284–285.

Odebrecht, Carolin (2015): „Interdisziplinäre Nutzung von Forschungsdaten mithilfe einer technisch-abstrakten Modellierung“, in: *DHd 2015: Von Daten zu Erkenntnissen* https://dh2014.files.wordpress.com/2014/07/

dh2014_abstracts_proceedings_07-11.pdf [letzter Zugriff 22. August 2016].

Odebrecht, Carolin / Krause, Thomas / Lüdeling, Anke (2015): „Austausch von historischen Texten verschiedener Sprachen über das LAUDATIO-Repository“, in: *DGfS-CL Poster Session. 37. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft* http://asvdoku.informatik.uni-leipzig.de/dgfs2015cl-ps/index.html [letzter Zugriff 22. August 2016].

Romary, Laurent / Ide, Nancy (2004): „International standard for a linguistic annotation framework“, in: *Natural Language Engineering* 10 (3-4): 211–225.

Romary, Laurent / Zeldes, Amir / Zipser, Florian (2015): „<tiger2/>: serialising the ISO SynAF syntactic object model“, in: *Language Resources and Evaluation* 49 (1): 1–18.

Rümpel, Stefanie (2011): „Der Lebenszyklus von Forschungsdaten“, in: Büttner, Stephan / Hobohm, Hans-Christoph / Müller, Lars (eds.): *Handbuch Forschungsdatenmanagement*. Bad Honnef: Bock + Herchen 25–34.

Schmidt, Thomas / Wörner, Kai (2009): „EXMARaLDA. Creating, analysing and sharing spoken language corpora for pragmatic research”, in: *Pragmatics* 19 (4): 565–582.

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): „Guidelines für das Tagging deutscher Textkorpora mit STTS“, in: Universität Tübingen (ed.): Seminar für Sprachwissenschaft. Technischer Report. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf [letzter Zugriff 22. August 2016].

Solodovnik, Iryna (2011): „Metadata issues in Digital Libraries. key concepts and perspectives“, in: *Italian Journal of Library, Archives and Information Science* 2 (2): 4663-1–4663-27. 10.4403/jlis.it-4663.

Stiller, Juliane / Thoden, Klaus / Zielke, Dennis (2016): „Usability in den Digital Humanities am Beispiel des LAUDATIO-Repositoriums“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 244–247.

TEI Consortium (2015): *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (2.9.1). http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 15. November 2015].

Van Uytvanck, Dieter / Stehouwer, Herman / Lempen, Lari (2012): ”Semantic metadata mapping in practice: the Virtual Language Observatory“, in: *Proceedings of LREC 2012* 1029–1033.

Vertan, Cristina / Ellwardt, Andreas / Hummerl, Susanne (2016): „Ein Mehrebenen-Tagging-Modell für die Annotation altäthiopischer Texte“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 258–261.

Xie, Iris / Matusiak, Krystyna (2016): *Discover Digital Libraries. Theory and Practice*. Oxford: Elsevier.

Index der Autorinnen und Autoren

Ergänzungen

Aufgrund eines technischen Problems sind im Abstractband einige Bilder nicht übernommen worden. Die folgenden Seiten enthalten die fehlenden Abbildungen sortiert nach Beitrag.

Panel:
Virtuelle Forschungsumgebung für objekt- und raumbezogene Forschung

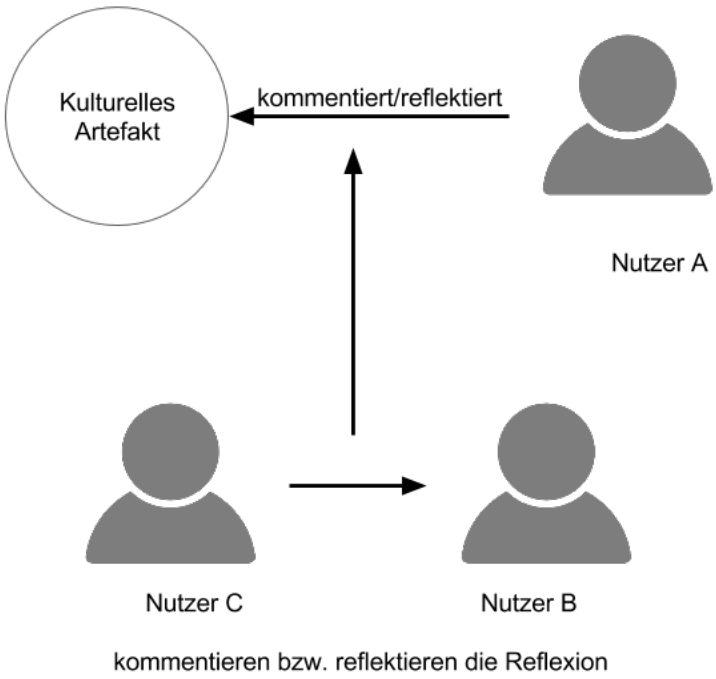
Kuroczyński, Piotr; Stanicka-Brzezicka, Ksenia; Fichtl, Barbara; Köhler, Werner; Brahaj, Armand; Fichtner, Mark

Abb. 1: Gesamtmenge an generierten Daten in den vergangenen Jahren (Quelle: http://edition.cnn.com/2014/11/04/tech/gallery/big-data- techonomics-graphs/)

Vortrag:
Nachhaltigkeit als Prozess: Zur konzeptionellen Funktion digitaler Technologien in der Nachhaltigkeitssicherung für historische Fotos im Projekt efoto-Hamburg

Schumacher, Mareike

Eine Verknüpfung der drei Ebenen könnte z.B. wie folgt aussehen:



Poster:
Comparison of Methods for Automatic Relation Extraction in German Novels

Krug, Markus; Wick, Christoph; Jannidis, Fotis; Reger, Isabella; Weimer, Lukas; Madarasz, Nathalie; Puppe, Frank

			Relation (binary)			4 types of Relations			all 57 types		
#relation instances in train/test						(1070 / 195)					
			F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall
Rule-based											
	3 core rules		68.0%	77.6%	60.5%	NA	NA	NA	NA	NA	NA
	all rules		71.0%	72.3%	69.7%	59.0%	74.2%	48.9%	NA	NA	NA
Feature-based											
	generic features										
		MaxEnt	67.3%	78.7%	58.7%	50.6%	69.0%	40.0%	41.2%	73.7%	28.6%
		SVM	58.0%	69.0%	50.0%	43.2%	56.7%	34.9%	42.1%	54.5%	34.4%
	generic + rule features										
		MaxEnt	73.6%	82.8%	66.3%	61.2%	79.5%	49.7%	52.8%	73.0%	41.3%
		SVM	69.2%	83.4%	59.2%	58.1%	78.3%	46.2%	45.1%	68.0%	33.7%

Tabelle 1: Ergebnisse der verschiedenen Ansätze für drei verschiedene Evaluationsszenarien: binär (das reine Vorliegen einer Relation), für die 4 Haupttypen und für alle 57 Relationstypen insgesamt.

			Family-relation			Love-relation			professional-relation			social-relation		
#relation instances in train/test			(508/93)			(260/38)			(88/18)			(203/45)		
			F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall
Rule-based														
	3 core rules		69.9%	96.2%	54.8%	NA	NA	NA	26.1%	60.0%	16.7%	NA	NA	NA
	all rules		78.0%	90.1%	68.8%	53.84%	52.5%	55.2%	19.0%	66.7%	11.1%	27.1%	57.1%	17.8%
Feature-based														
	generic features													
		MaxEnt	69.4%	76.6%	63.4%	38.5%	71.4%	26.3%	17.4%	50.0%	10.5%	22.2%	38.9%	15.6%
		SVM	60.1%	61.1%	59.1%	20.8%	50.0%	13.2%	10.0%	50.0%	5.3%	22.2%	38.9%	15.6%
	generic + rule features													
		MaxEnt	78.5%	91.4%	68.8%	56.3%	60.6%	52.6%	32.0%	66.7%	21.1%	31.0%	69.2%	20.0%
		SVM	77.7%	95.3%	65.6%	50.8%	71.4%	39.5%	26.0%	75.0%	15.8%	31.0%	42.3%	24.4%

Tabelle 2: Ergebnisse für die verschiedenen Ansätze, aufgeschlüsselt nach den 4 Haupttypen. Familienrelationen erreichen sehr gute Ergebnisse mit einem F1-Wert von fast 80% und einer Precision von bis zu 95%. Liebesrelationen sind schwerer zu erkennen, liegen aber dennoch bei 56,3% F1. Die anderen Relationstypen fallen in der Qualität ab, sind aber gleichzeitig weniger relevant.

Poster:
Digitale Erschließung einer Sammlung von Volksliedern aus dem deutschsprachigen Raum

Burghardt, Manuel; Spanner, Sebastian; Schmidt, Thomas; Fuchs, Florian; Buchhop, Katia; Nickl, Miriam; Wolff, Christian

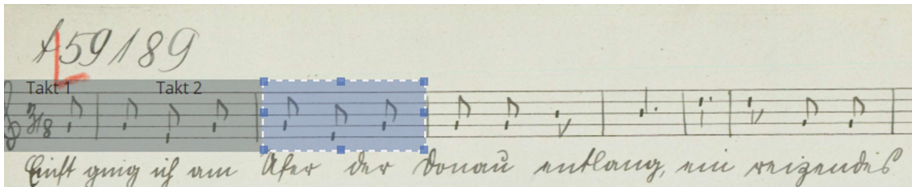


Abbildung 5: Taktweise Segmentierung der Liedblätter mit dem *Allegro*.