handbuch.io

- CH

Search

Search Go

- Views
- Actions
- Tools

DH-Handbuch/ Druckversion

From Handbuch.io

< DH-Handbuch

Error creating thumbnail: File missing

Die Version 1.0 ist auch als gelayoutetes Buch im PDF-Format verfügbar: http://bit.do/DH-Handbuch (PDF, 5MB)

Contents

Einführung: Projekte und Forschungsfragen in den Digital Humanities

Über dieses Handbuch

Das vorliegende Handbuch ist im Rahmen eines Book Sprints an der Open Knowledge Foundation[1] im August 2015 in Berlin entstanden. Ziel dieses Buchs ist konzentrierten Überblick über das Feld der *Digital Humanities* (DH) anzubieten. Für Einsteiger und mögliche AntragstellerInnen stellen sich häufig die folgend

- Was sind die Digital Humanities?
- Was sind relevante Forschungsfragen?
- Mithilfe welcher Tools lassen sich fachspezifische, aber auch fächerübergreifende Fragen beantworten?
- Was müssen Geisteswissenschaftler beim Umgang mit Daten beachten?
- Wie sehen erfolgreiche Projekte in den Digital Humanities aus?

Neben Lösungswegen und Ressourcen zu typischen Fragen werden auch Projekte und Werkzeuge detailliert vorgestellt, um vorhandene Kenntnisse aufzufrische Aspekte der Digital Humanities kennenzulernen. Die Nähe zur fachwissenschaftlichen Praxis steht dabei im Vordergrund. Wir hoffen, mit diesem Handbuch auc Digital Humanities nahebringen zu können und die Neugierde auf digitale Methoden und deren Möglichkeiten für die geisteswissenschaftliche Forschung zu we

Was sind die Digital Humanities?

Bestimmte Forschungsfragen lassen sich durch den Einsatz von Computern besser beantworten als mithilfe konventioneller, nicht-digitaler Methoden der Geiste Andere geisteswissenschaftliche Fragen lassen sich überhaupt nur bearbeiten, weil es digitale Methoden und Verfahren gibt. Ob digitale Methoden und Verfahre werden sollten, hängt dabei wesentlich von den Forschungsfragen ab, die im Zentrum des geisteswissenschaftlichen Interesses stehen.

So empfiehlt sich der Einsatz digitaler Werkzeuge insbesondere dann, wenn sehr große Datenmengen untersucht, sehr lange Perioden fokussiert, oder feinste Un zwischen Inhalten erkannt werden sollen.

Die Vorteile des Computereinsatzes sind bekannt: Maschinen ermüden nicht, erkennen Muster ohne Erwartungen und verzählen sich nicht. Dennoch ist die Wah Methoden als auch die Erstellung oder Auswahl von Korpora (also der Datengrundlage) eine intellektuell anspruchsvolle Aufgabe, die erfahrene Wissenschaftle Anforderungen ihrer Forschungsfragen durchführen sollten. Die Interpretation der Ergebnisse computergestützter Analysen setzt ein breites Verständnis der eing Methoden voraus und sollte sich auch mit den Grenzen digitaler Methoden auseinandersetzen, sowie die verwendete Datenbasis kritisch hinterfragen.

Digital Humanities finden genau in diesem Spannungsfeld zwischen geisteswissenschaftlichen Fragestellungen, traditionellen Quellen und den Möglichkeiten v Werkzeugen statt. Dabei wurde schon viel über die erwünschte Ausbildung von digitalen GeisteswissenschaftlerInnen, ihre Arbeitsweise und Schnittstellen zu tr Disziplinfeldern geschrieben.[2]

Die angesprochenen Diskussionen gelangen zu ganz unterschiedlichen Schlussfolgerungen: So wird teilweise antizipiert, dass bereits alle Geisteswissenschaft d und es daher keinen definitorischen Bedarf gibt, andererseits wird von traditionellen Vertretern ihres Fachs eine "feindliche Übernahme"[3] durch die Informatik darin resultiert, dass alle hermeneutische geisteswissenschaftliche Arbeitsweise nicht mehr genügt.

Die Autoren dieses Bandes werden diese definitorischen Probleme nicht lösen können. Wir stellen aber fest: Es gibt die Digital Humanities. Im weitesten Sinne dabei um die Beantwortung geisteswissenschaftlicher Fragestellungen mithilfe digitaler Methoden. Der Einsatz von Office Programmen fällt darunter ebenso w Verwendung von Wikis oder E-Mail. Diese Tools unterstützen lediglich die Kommunikation, erleichtern das wissenschaftliche Schreiben und dienen selbst tradi Vertretern des Fachs als alltägliche Werkzeuge.

Es existiert aber eine vielfältige Reihe von Forschungsfragen aus den Geisteswissenschaften, die sich mit Hilfe der Digital Humanities elegant beantworten lasse Liste bietet eine Reihe interessanter Ansätze an, erhebt aber keinen Anspruch auf Vollständigkeit.

Forschungsfragen

Auf die Frage von Gregory Crane “What do you do with a million books?”[4] antworten Clement et al.

> „You don't read them, because you can't“
> – Clement, Steger und Unsworth, Kirsten Uszkalo: How Not to Read a Million Books, 2008

und bieten im Folgenden zahlreiche Fragen, die sich nur im Massenzugriff auf strukturierte Textdaten beantworten lassen, zum Beispiel

> „Words that Jane Austen uses less often than other novelists 1780-1830.“
> – Clement, Steger und Unsworth, Kirsten Uszkalo: How Not to Read a Million Books, 2008, http://people.brandeis.edu/~unsworth/hownot2read.html

Ein Beispiel für eine Methode der Digital Humanities ist das irreführender Weise "distant reading" – besser Makroanalyse[5] – genannte Verfahren. Aufgrund de Menge von Büchern, die jeder Mensch in seinem Leben lesen kann [6] und der gleichzeitigen Neugier auf die Inhalte vieler weiterer Bücher, ist die einzige Mög Informationen aus weiteren Büchern zu verarbeiten, deren automatische, d.h. algorithmische Durchdringung und Aufarbeitung. Dies wird mithilfe zahlreicher M Einsatz verschiedener Software und Algorithmen – bewerkstelligt.

Ein konkretes Beispiel für solche Forschungsprojekte ist die quantitative Textanalyse. Hier wurden beispielsweise im Rahmen der Analyse antiker ägyptischer T interessante Ergebnisse erzielt.[7] Auch auf die Stimmungen von Epochen lassen sich solche Analysen anwenden: "Roaring Twenties", "Les Trente Glorieuses" (auch "Nach dem Boom" (ab den 1970er Jahren) sind allgemein akzeptierte Charakterisierungen von mit stark positiven oder negativen Emotionen besetzten Epo Muster finden wir auch in den Texten aus diesen Epochen[8].

Weiterführende Untersuchungen könnten das Verhältnis zwischen Tagespresse und der Belletristik untersuchen: In welcher Textgattung kündigen sich Stimmung früher an? Inwieweit handelt es sich um globale Phänomene, oder zeigen sich je nach Region oder Sprache zeitlich verschobene Stimmungsphasen? Sind deutsc – wie es der Begriff der "German Angst" vermuten ließe – wirklich emotional anders geprägt als englischsprachige? Und falls ja, gibt es Zeiten besonders starke auch der Konvergenz zwischen diesen beiden Sprachräumen?

Ähnliche Verfahren lassen sich auf andere Medien übertragen, hier besteht die Möglichkeit durch maschinelle Mustererkennung regelmäßig wiederkehrende Ele Bilddateien zu erkennen und miteinander zu vergleichen.

Zur Frage "How to compare 1 Mio Images?" wurde beispielsweise eine Studie publiziert[9], in der diverse Projekte aufgeführt werden. Auch die Möglichkeiten Technologie sind für Disziplinen, welche sich mit Gegenständen im Raum (beispielsweise Kunstgeschichte oder Archäologie) beschäftigen, von großem Interes man sowohl nutzbar machen, um öffentliche Räume einer Epoche nachzubilden und daraus Schlüsse über gesellschaftliche Belange aus besagter Epoche zu zieh archäologische Gegenstände zu digitalisieren und auf dieser Datenbasis mithilfe algorithmischer Verfahren maschinelle Musterkennung zu betreiben.[10]

Daneben bieten Methoden der sozialen Netzwerkanalyse interessante Möglichkeiten für verschiedene Geisteswissenschaften, beispielsweise die Geschichtswiss können Relationen zwischen Personen oder Personengruppen in vergangenen Gesellschaften mithilfe von Methoden der sozialen Netzwerkanalyse untersucht w

Error creating thumbnail: File missing

Beispiel für historische Netzwerkanalyse. Visualisierung
tausender Dokumente, die zwischen Völkerbund Experten
während der Zeit zwischen den Weltkriegen ausgetauscht
wurden. Von Martin Grandjean, Quelle:
https://en.wikipedia.org/wiki/Digital_history#/media/File:Social_Network_Analysis_Visualization.png.
CC BY-SA 3.0

Weitere Forschungsfragen, die mit Methoden der Digital Humanities beantwortet werden könnten:

- Was kann man mit einem Korpus von hunderten Inkunabel-Abbildungen machen? Gemeinsamkeiten und Unterschiede zwischen den abgebildeten Person es immer wieder kehrende Personen? In welchen Farbtönen / mit welchen Gegenständen werden diese abgebildet? die Beziehung zwischen Hauptakteure (durch Mustererkennung in Bildern? bzw. Messung des durchschnittlichen Abstandes?)
- Im Nachlass einer Autorin finden sich hunderte von Gedichten, welche jeweils dutzende unterschiedliche Fassungen haben. Wie lässt sich dieses Werk in abbilden? Lassen sich die Fassungen in eine chronologische Reihenfolge bringen?
- Bekannte Vertreter einer Epoche sind durch regelmäßige Briefkontakte verbunden. Wie bildet man dieses Netzwerk vollständig ab und wie lässt es sich ar
- Eine Erzählung erscheint anonym und mehrere bekannte Autoren kämen als VerfasserInnen in Frage. Wie lässt sich die Herkunft korpusbasiert untersuche
- Zahlreiche Texte einer Epoche beziehen sich auf bestimmte geographische Orte. Wie lassen sich die Schnittmengen dieser Bezüge abbilden und analysier
- Welche Implikationen bergen Computerspiele in Hinblick auf die Simulation historischer Gesellschaften?
- Welche Fragen kann man mit Methoden der künstlichen Intelligenz beantworten?

Aufbau des Handbuchs

Um das Handbuch möglichst praxisnah zu gestalten, haben wir uns entschieden, zuerst einzelne DH-Projekte vorzustellen, um die Möglichkeiten der DH den Le bringen und ihnen zu zeigen, was in der Praxis in dem Bereich derzeit schon umgesetzt wurde. So zeigen wir in Kapitel 2, wie mit TextGrid Texte editiert und m eCodicology Handschriften analysiert werden. Die folgenden drei Kapitel beschäftigen sich mit den drei Säulen, die jedes Projekt in den Digital Humanities trag Methoden und Werkzeuge, und Infrastruktur. Die Kapitel bieten erste Einführungen in die jeweilige Thematik und vermitteln den Lesern an die Praxis angelehnt sie in eigenen DH-Projekten anwenden können. Die Kapitel Daten und Alles was Recht ist - Urheberrecht und Lizenzierung von Forschungsdaten weisen in die Grundlage wissenschaftlichen Forschens ein und bieten Hilfestellungen im Umgang mit Lizenzen und Dateiformaten. Das Kapitel Methoden und Werkzeuge ze Digital Humanities auf und verweist beispielhaft auf digitale Werkzeuge, die für die Beantwortung geisteswissenschaftlicher Forschungsfragen herangezogen w Kapitel Infrastruktur werden Digitale Infrastrukturen, deren Komponenten und Zielstellungen näher beschrieben. Sie sind unerlässlich, um die digitale Forschun und nachhaltig zu gestalten.

Autoren

Alle Autorinnen und Autoren in alphabetischer Reihenfolge:

```
Helene Hahn
Tibor Kalman
Steffen Pielström
Johanna Puhl
Wibke Kolbmann
Thomas Kollatz
Markus Neuschäfer
Juliane Stiller
Danah Tonne
```

Folgende Personen sind mit Interviewbeiträgen vertreten:

```
Anne Baillot
Peter Bell
Daniel Burckhardt
Hannah Busch
Matthias Kiesselbach
Christoph Kümmel
Thomas Kollatz
Andrea Rapp
Patrick Sahle
```

Links und Literatur

Eine annotierte Bibliografie zu den Digital Humanities wird fortlaufend in DARIAH-DE geführt: Doing Digital Humanities https://de.dariah.eu/bibliographie

Einführungen in die Digitalen Geisteswissenschaften / Digital Humanities gibt es reichlich, hier eine (kleine) Auswahl:

Johanna Drucker. 2014. *Intro to Digital Humanities. Concepts, Methods, and Tutorials for Students and Instructors.* Online course book: http://dh101.humanitie

Willard McCarty. 2005. *Humanities Computing.* Basingstoke & New York: Palgrave Macmillan. http://www.mccarty.org.uk/essays/McCarty,%20Humanities%2

Ray Siemens, John Unsworth, and Susan Schreibman. 2004. *A Companion to Digital Humanities*. Blackwell Companions to Literature and Culture. http://www.digitalhumanities.org/companion/

Susanne Kurz. 2014. *Digital Humanities: Grundlagen und Technologien für die Praxis.* Springer. ISBN: 978-3658057923

Forschungsdaten sind z.B. unter den folgenden Quellen zu finden:

https://www.openaire.eu/search/find

Anmerkungen

1. ↑ Open Knowledge Foundation Deutschland e.V. - http://okfn.de
2. ↑ https://dev2.dariah.eu/wiki/download/attachments/14651583/DARIAH-M2-3-3_DH-programs_1_1.pdf?version=2&modificationDate=1366904376117
3. ↑ Vgl. http://www.hsozkult.de/conferencereport/id/tagungsberichte-5384
4. ↑ Gregory Crane: What Do You Do with a Million Books?, D-Lib Magazine 12 (2006), http://www.dlib.org/dlib/march06/crane/03crane.html
5. ↑ Jockers, Matthew L.: Macroanalysis: Digital Methods and Literary History (University of Illinois Press, 2013)
6. ↑ Vgl. http://www.bookpedia.de/buecher/Wieviel_kann_ein_Mensch_in_seinem_Leben_lesen%3F
7. ↑ http://totenbuch.awk.nrw.de/projekt/das-totenbuch
8. ↑ http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0059030
9. ↑ http://softwarestudies.com/cultural_analytics/2011.How_To_Compare_One_Million_Images.pdf
10. ↑ http://vpcp.chass.ncsu.edu/, http://romereborn.frischerconsulting.com/gallery-current.php, http://www.forumromanum30.hu-berlin.de/, http://www.educause.edu/ero/article/virtual-paul%E2%80%99s-cross-project-digital-modeling%E2%80%99s-uneasy-approximations.
11. ↑ Lemercier, Claire. “Formale Methoden Der Netzwerkanalyse in Den Geschichtswissenschaften: Warum Und Wie?” In *Historische Netzwerkanalyse*. Inr Verlag, n.d.

Digital Humanities in der Praxis

Wie verlaufen DH-Projekte? Einblicke in die Praxis

Um einen Einblick in die tagtägliche Arbeit von WissenschaftlerInnen in den Digital Humanities zu bekommen, haben wir einige WissenschaftlerInnen gebeten, vorzustellen und ihrer digitalen Methoden und erzielte Ergebnisse näher zu beleuchten.

Anne Baillot: Berliner Intellektuelle 1800-1830

Anne Baillot ist Literaturwissenschaftlerin und leitet seit 2010 am Institut für deutsche Literatur der Humboldt-Universität zu Berlin eine Emmy Noether-Nachw
Thema "Berliner Intellektuelle 1800-1830". Sie koordiniert den Einstein-Zirkel Digital Humanities.

Error creating thumbnail: File missing

Briefe und Texte aus dem intellektuellen Berlin um 1800:
http://tei.ibi.hu-berlin.de/berliner-intellektuelle/

Wie ist die Idee zu dem Projekt entstanden?

Anne Baillot: Das Projekt "Berliner Intellektuelle 1800-1830" schloss direkt an meine Dissertation an. In meiner Dissertation habe ich die Frage nach der Konsti
Intellektuellennetzwerken im Kontext der napoleonischen Kriege an einem Einzelbeispiel untersucht und wollte die Analyse ausdehnen, um ein strukturelles Ver
Netzwerke zu gewinnen und dieses in der Lektüre der produzierten (literarischen, geisteswissenschaftlichen) Texte umzusetzen. Der Skalawechsel vom Einzelfa
Gesamtstruktur war eine Grundvoraussetzung des Projektes.

Wie lautete die Fragestellung zu dem Projekt?

Anne Baillot: Im Mittelpunkt des Projektes steht die Frage nach Form und Bedeutung der Teilnahme von Gelehrten am öffentlichen Leben mit besonderer Berüc
Kommunikationsstrategien und der damit einhergehenden politischen Stellungnahmen. Untersucht werden die Berliner Intellektuellennetzwerke zwischen 1800
des Kultur- und Wissenstransfers.

Warum wurden digitale Methoden gewählt?

Anne Baillot: Die Grundidee bestand darin, mehrere Textkorpora (die jeweils einen der einschlägigen thematischen Schwerpunkte der Fragestellung illustrierten
gemeinsamen Rasters zu erfassen. Dies war viel ökonomischer und auf einer viel größeren Skala durch digitale Mittel zu bewerkstelligen. So entstand die digita
und Texte aus dem intellektuellen Berlin um 1800" (http://tei.ibi.hu-berlin.de/berliner-intellektuelle/ -NB: Der Techniker macht diese Tage größere Updates, es k
Edition mal nicht geht)

Wie wurden die Daten erhoben?

Anne Baillot: Die Handschriften wurden nach ihrer Relevanz für die Fragestellung ausgesucht. Die Transkription und Annotation erfolgte komplett händisch, au
biographischen Datensätze, die im Rahmen des Boeckh-Nachlassprojektes (http://tei.ibi.hu-berlin.de/boeckh/) über die GND-Nummern von kalliope
(http://kalliope.staatsbibliothek-berlin.de/de/index.html) direkt importiert wurden.

Welche Tools haben Sie ausgewählt und warum?

Anne Baillot: Die genaue Struktur der Datenbank kenne ich nicht. Am Ende der Literaturwissenchaftler wird in XML/TEI gearbeitet, mit Oxygen. Die Einarbeit
oxygen sehr gut. Textgrid wurde zu Projektbeginn einstimmig verworfen, da zu unstabil. In oxygen ist ein SVN-Client eingebaut, der die kollaborative Anreiche
möglich macht.

Wie verlief die Analyse?

Anne Baillot: Die Analyse erfolgte soweit primär analog. In Zusammenarbeit mit Machine Learning-Spezialisten der Technischen Universität arbeite ich derzeit
großformatigeren Auswertung bestimmter Textphänomenen, die von uns annotiert wurden. Darüber hinaus hoffe ich, Netzwerkanalyse anschliessen zu können,
meiner Sicht nur in Zusammenarbeit mit Informatikern erfolgen kann (und es hoffentlich wird).

Wie wurden die Ergebnisse publiziert?

Anne Baillot: Digitale Edition, Nachlassverzeichnis, Blog, Sammelbände, Aufsätze wurden bereits veröffentlicht. Darüber hinaus sind 3 Dissertationen in der Al

Weitere Links:

http://digitalintellectuals.hypotheses.org/

http://www.literaturkritik.de/public/rezension.php?rez_id=19678&ausgabe=201409

Daniel Burckhardt: Verbrannte und Verbannte

Daniel Burckhardt ist Mathematiker und Wissenschaftshistoriker an der Humboldt-Universität zu Berlin. Zusammen mit einem ad-hoc Team von 9 Personen set
Projekt "Verbannte und Verbrannte" um, eine Webseite zur Liste der im Nationalsozialismus verbotenen Publikationen und Autoren.

Error creating thumbnail: File missing

Visualisierung von Lebenswegen auf der Seite "Verbrannte und Verbannte": http://verbrannte-und-verbannte.de/

Wie ist die Idee zu dem Projekt entstanden?

Daniel Burckhardt: Das Projekt entstand im Rahmen von {COD1NG DA V1NC1}, dem ersten „Kultur-Hackathon“ in Berlin, der zwischen dem 26./27. April u 2014 stattfand.

Wie lautete die Fragestellung zu dem Projekt?

Daniel Burckhardt: Das Projekt startete weniger mit einer Fragestellung als mit einem Datensatz, der vom Land Berlin als offene Daten veröffentlichte Liste der Bücher. Dieser Basisdatensatz wurde in einem zweiten Schritt systematisch mit Normdaten ergänzt. Ziel war einerseits eine bessere Nutzerführung, andererseits inhaltlich ergänzten Daten die Basis für Visualisierungen und statistische Auswertungen.

Warum wurden digitale Methoden gewählt?

Daniel Burckhardt: Eine manuelle Bearbeitung der rund 5'000 bibliografischen Einträge sowie fast 2'000 Personendatensätze wäre im kurzen Zeitraum von gut 7 zu leisten gewesen.

Wie wurden die Daten erhoben?

Daniel Burckhardt: Der Basisdatensatz wurde vom Land Berlin zur Verfügung gestellt. Diese Daten wurden mit den Katalogdaten der Deutschen Nationalbiblio Format abgeglichen. Da die Verfasserinnen, Herausgeber und Verlage im Katalog mit GND-Nummern markiert sind, konnten automatisiert Zusatzinformationer Linked-Open-Data-Diensten (Entity Facts, Wikidata) abgerufen werden.

Welche Tools haben Sie ausgewählt und warum?

Daniel Burckhardt: OpenRefine zum Bereinigen der Daten, Programmcode in Java und PHP (https://github.com/jlewis91/codingdavinci, https://github.com/mhinters/BannedBookUtils), JavaScript-Bibliotheken (Leaflet.js, D3.js) zur Präsentation.

Wie verlief die Analyse?

Daniel Burckhardt: Im wesentlichen wurden die Daten über SQL-Abfragen nach verschiedenen Kriterien gruppiert und dann auf Karten oder als Diagramme vis

Wie wurden die Ergebnisse publiziert?

Daniel Burckhardt: Bislang nur über die Website. Eine fachwissenschaftliche Analyse der Ergebnisse steht noch aus.

Weitere Informationen

Verbrannte und Verbannte. Die Liste der im Nationalsozialismus verbotenen Publikationen und Autoren: http://verbrannte-und-verbannte.de/about

Matthias Kiesselbach und Christoph Kümmel (DFG): Digital Humanities aus Förderperspektive

Error creating thumbnail: File missing

http://www.dfg.de/

Matthias Kiesselbach ist in der DFG-Geschäftsstelle zuständig für das Fach der Philosophie und in der Gruppe Geistes- und Sozialwissenschaften Ansprechpart Digitalisierung in der geistes- und sozialwissenschaftlichen Forschung.

Christoph Kümmel ist in der Geschäftsstelle der DFG zuständig für das Förderprogramm Fachinformationsdienste für die Wissenschaft sowie für das „Bilateral Humanities Program“ mit dem National Endowment for the Humanities (NEH). In der Gruppe Wissenschaftliche Literaturversorgungs- und Informationssystem Ansprechpartner zu Fragen der digitalen Informationsinfrastrukturen für die Geisteswissenschaften.

Wie schätzen Sie die Entwicklung der digitalen Geisteswissenschaften ein?

In den letzten Jahren wurden an mehreren Universitäten Zentren für die Forschung und Lehre in den Digitalen Geisteswissenschaften gegründet; 2013 kam ein I hinzu (DHd – Digital Humanities im deutschsprachigen Raum); im letzten Wissenschaftsjahr des BMBF (Motto: „Digitale Gesellschaft“) wurden die Digitalen Geisteswissenschaften immer wieder an prominenter Stelle erwähnt. Angesichts dieser Entwicklungen kann man mit Fug und Recht behaupten, dass die digitale Geisteswissenschaften eine Kraft sind, mit der man in der Wissenschaftsszene rechnen muss.

Allerdings bleibt zweierlei zu konstatieren. Zum einen ist das Aufkommen der „digitalen Geisteswissenschaften“ nicht unbedingt ein qualitatives Novum. In Be und Methoden waren verschiedene Bereiche der Geisteswissenschaften schon immer offen und pragmatisch. Etwa für die Archäologie oder die Sprachwissensch bestimmte digitale Technologien schon lange zum Quellen- und Methodenportfolio, bevor sich die Rede von den „digitalen Geisteswissenschaften“ etablierte. Z großer Teil der Anstrengungen im Bereich der sogenannten Digitalen Geisteswissenschaften bislang eher auf den forschungsvorbereitenden Bereich als auf die F konzentriert. Mitunter ist in den Foren und Veranstaltungen der Digitalen Geisteswissenschaften mehr von digitalen Werkzeugen als von konkreten Forschungsf zu deren Beantwortung die Werkzeuge dienen sollen.

Andererseits ist die Dynamik im Feld der der digitalen Geisteswissenschaften unbestreitbar. Die große Frage ist, was in der Zukunft von den digitalen Geisteswi erwarten ist. Werden die digitalen Technologien sich weiterhin (vor allem) einfügen in das Quellen- und Methodenportfolio der Geisteswissenschaften und („blo dass größere Datenmengen schneller analysiert werden können oder dass die Validität der Ergebnisse steigt? (Das wäre gewiss nicht wenig!) Oder werden sie, w der digitalen Geisteswissenschaften nicht selten anzukündigen scheint, die Disziplinen in einem substanzielleren Sinn transformieren?

Diese Frage ist aus Sicht der DFG noch weitgehend offen. Ihre Antwort wird – auch – von der Qualität und Ernsthaftigkeit der Selbstreflexion der Digitalen Geisteswissenschaften abhängen. Dazu zählt die noch kaum begonnene Diskussion der epistemologischen Grundlagen der Geisteswissenschaften im Licht ihrer Erweiterungen, und hier insbesondere die Frage, wie sich der im Kern hermeneutische Ansatz der klassischen Geisteswissenschaften eigentlich zu den Möglichk digitalen Technologien verhält. Vermutlich wird sich der Beitrag der Digitalität zur geisteswissenschaftlichen Forschung irgendwo zwischen »more of the same« but better«) und der Transformation der klassischen Forschungspraxis durch den Beitrag des Digitalen liegen. Aber wie er genau aussieht, bleibt abzuwarten.

Welche Art von Digital Humanities-Projekten fördert die DFG?

Grundsätzlich gibt es zwei Arten von DH-Projekten, die von der DFG gefördert werden können. Zum einen sind dies Projekte, die unter den Oberbegriff der Infrastrukturentwicklung fallen. Dabei geht es beispielsweise um die digitale Aufbereitung von Quellen – etwa die Retrodigitalisierung gedruckter Texte in Bibl Ermöglichung neuer Forschungsperspektiven oder um die Entwicklung und Erprobung neuartiger Werkzeuge, Organisationsformen und „Geschäftsmodelle“ für

forschungsrelevanten Informationen – nicht zuletzt für Publikationen. Dabei ist zu bedenken, dass auch in diesem Bereich grundsätzlich nur Projekte gefördert v also keine Daueraufgaben von Informationseinrichtungen querfinanziert werden können.

Zum andern werden Forschungsprojekte im engeren Sinn gefördert. Dies können Projekte aller Art und jeglicher Dimension sein. Für Projekte der Digitalen Gei gibt es keine eigenen Regeln. Hier gilt wie überall bei der DFG: Solange es sich bei einem Projekt um einen innovativen und vielversprechenden wissenschaftlic handelt, kann es von der DFG gefördert werden. Und wenn bei einem Projekt die Einbeziehung digitaler Technologien oder informatischer Expertise notwendig wissenschaftlichen Projekterfolg ist – dann wird auch sie gefördert. Die Einschätzung der Frage, inwieweit diese Kriterien bei einem konkreten Projekt erfüllt si externen Gutachter*innen und der Entscheidungsgremien.

Welche Zeiträume und Fristen sind bei Neuanträgen zu beachten?

Bei der DFG gibt es prinzipiell keine Einreichungsfristen für Projektanträge.

Ausnahmen bilden (bestimmte) internationale Ausschreibungen, darunter auch die in Kooperation mit dem amerikanischen National Endowment for the Human bislang alle zwei Jahre stattfindende Digital Humanities-Ausschreibung. Eine weitere Ausschreibung mit fester Einreichungsfrist ist die in Kooperation mit vers europäischen und nord- sowie südamerikanischen Förderorganisationen geplante Initiative „Digging Into Data“. Die Ausschreibung wird voraussichtlich im Wi für deutsche Teilnehmer*innen geöffnet.

In der Softwareentwicklung ist es häufig sinnvoll, digitale Werkzeuge in verschiedenen Iterationen zu entwickeln (agile Softwareentwicklung), um sich Tests mit einer vorläufigen Version auf neue Features zu einigen. Wie lässt sich dies mit den Anforderungen der Anträge vereinbaren?

Die Frage ist für uns nicht auf Anhieb verständlich – wahrscheinlich ist damit angedeutet, dass bei moderner Softwareentwicklung „Weichen“ und „Entscheidun eingeplant werden müssen und sich der Erfolg nicht im Voraus planen lässt. Es gibt die Beobachtung, dass Anträge, bei denen es um Softwareentwicklung geht Arbeitsplan in keiner Weise auf Risiken, alternative Szenarien oder bewusst eingebaute Tests (mit Konsequenzen für den Fortgang) eingegangen wird, von den skeptisch beurteilt werden. Grundsätzlich muss man davon ausgehen, dass die ausgewählten Gutachter*innen Experten ihres Faches sind und sich bestens damit Softwareentwicklung zu einem Erfolg werden kann.

Wo kann man sich zu weiteren Fördermöglichkeiten außerhalb der DFG informieren?

Sicherlich auf dem Webauftritt des DHd. Lohnend ist immer auch ein Blick auf die Seiten des BMBF und anderer einschlägiger Stiftungen (beispielsweise der VolkswagenStiftung).

Welche Bedeutung hat die Nachnutzbarkeit von Forschungsdaten?

Die DFG legt großen Wert auf die freie Zugänglichkeit und Nachnutzbarkeit von Forschungsdaten. Von allen Antragsteller*innen, die Projekte planen, innerhalb Forschungsdaten generiert werden, wird eine ernsthafte und sorgfältig dokumentierte Beschäftigung mit den Nachnutzungsmöglichkeiten der generierten Daten Maßnahmen zur Ermöglichung der Nachnutzung von Forschungsdaten im Einzelnen erwartet werden, ist allerdings nicht allgemein zu sagen. Es gilt, dass der zu Aufwand für eine konkrete Maßnahme in einem sinnvollen Verhältnis zum erwarteten Nutzen stehen muss; ebenfalls erkennt die DFG an, dass in verschiedenen unterschiedliche Standards gelten. Einschränkend muss man auch leider feststellen, dass es in manchen Wissenschaftsbereichen noch keine verlässlichen Infrast entsprechende Angebote machen. Aber die Lage bessert sich!

Weitere Informationen zur Nachnutzung von Forschungsdaten lassen sich unter
http://www.dfg.de/foerderung/antragstellung_begutachtung_entscheidung/antragstellende/antragstellung/nachnutzung_forschungsdaten/index.html abrufen.

Wie wichtig ist Open Access in den Geisteswissenschaften?

Ebenso wichtig wie die Nachnutzbarkeit von Forschungsdaten ist die ungehinderte Verbreitung von Forschungsergebnissen. Aus diesem Grund fördert die DFG Bewegung auf vielfältige Weisen und fordert auch ihre Drittmittelempfänger auf, bei der Veröffentlichung ihrer Forschungsergebnisse nach Möglichkeit auf Ope Publikationsorte zu setzen. Die DFG erkennt allerdings an, dass die verfügbaren Open Access Publikationsorte nicht in allen Fächern die üblichen Standards der Qualitätssicherung erfüllen. Wir erleben hier in den letzten Jahren jedoch eine enorme Zunahme an sehr guten Publikationsmöglichkeiten – auch in Fächern, die ganz auf Printpublikationen gesetzt haben. Wir gehen davon aus, dass es zunehmend unproblematischer wird, eine – auch aus engerer fachlicher Sicht – geeigne eine qualitativ hochwertige Veröffentlichung im Open Access zu finden. Künftig werden beispielsweise auch die „Fachinformationsdienste für die Wissenschaft Dienste anbieten.

Was können Forschungsinfrastrukturen zu DH-Projekten beitragen?

DH-Projekte – wie auch immer diese definiert sein mögen – profitieren genauso wie alle anderen Forschungsvorhaben von einer möglichst verlässlichen und tec Infrastruktur. Es ergibt sich aus der Natur der Sache, dass Projekte, die große Datenmengen analysieren möchten und hierfür beispielsweise auf entsprechende d Quellensammlungen („Corpora“) und geeignete Werkzeuge zum Umgang mit diesen Daten angewiesen sind, im besonderen Maße von neuartigen Infrastrukture viele DH-Projekte wäre es beispielsweise beruhigend zu wissen, wie es mit den großen Projekten auf der ESFRI-Roadmap (DARIAH, CLARIN) weitergeht.

Wo finden sich hilfreiche Links und Ressourcen zur Antragstellung?

Auf der Homepage der DFG finden Antragsteller*innen alles, was sie an Informationen für die Antragstellung brauchen.

Links

http://www.dfg.de

Die Deutsche Forschungsgemeinschaft ist die Selbstverwaltungsorganisation der Wissenschaft in Deutschland. Sie dient der Wissenschaft in allen ihren Zweige die DFG als privatrechtlicher Verein. Ihre Mitglieder sind forschungsintensive Hochschulen, außeruniversitäre Forschungseinrichtungen, wissenschaftliche Ver Akademien der Wissenschaften.

Die DFG erhält ihre finanziellen Mittel zum größten Teil von Bund und Ländern, die in allen Bewilligungsgremien vertreten sind. Dabei stellen Stimmverhältnis Verfahrensregeln wissenschaftsgeleitete Entscheidungen sicher.

Thomas Kollatz: Relationen im Raum visualisieren mit dem Topographie-Visualisierer

Thomas Kollatz ist wissenschaftlicher Mitarbeiter am Steinheim-Institut für deutsch-jüdische Geschichte in Essen. Seit 2002 entwickelt und betreut er epidat, di Datenbank zu historischen jüdischen Friedhöfen.

Wie ist die Idee entstanden?

Error creating thumbnail: File missing

Topographie Visualisierer

Thomas Kollatz: Die Idee zum Projekt "Relationen im Raum – Visualisierung topographischer Klein(st)strukturen" ist aus unserer jahrelangen epigraphischen A historischen jüdischen Friedhöfen entstanden. In den letzen Jahren wurde ein umfangreicher Inschriftenbestand auch digital erfasst. Zudem hatten wir zu einiger schematische Lagepläne. Die Projektidee war die Fülle an zu einem Objekt gesammelten Einzelinformation in einem zweiten Schritt wieder auf die Fläche zu br Einzelobjekt (Grabmal) also in Relation zum räumlichen Ensemble (Friedhof) zu setzen, um auf diese Weise möglicherweise Muster und Relationen zwischen d in den Blick zu bekommen. Zudem wurde neben der philologisch-historischen, textorientierten Perspektive auf die Grabmale, ein Partner gefunden, der die Obje bauhistorisch-kunstwissenschaftlicher Perspektive untersucht. Geplant war, einen Topographie-Visualisierer zu entwickeln, mit dem sich beliebige Phänomene g Suchinterface auf dem interaktiven Lageplan darstellen und analysieren lassen.

Wie wurde das Projekt finanziert?

Thomas Kollatz: Gefördert wird das Projekt vom BMBF im Rahmen der Förderlinie eHumanities (Förderkennzeichen: 01UG1243A-C).

Mit welchen Partnern wurde das Vorhaben umgesetzt und wie waren die Rollen verteilt?

Thomas Kollatz: Projektpartner sind:

- Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte, Essen

Das Steinheim-Institut vertritt im Verbund die epigraphisch, philologisch-historischen Fragestellungen.

- Bau- und Stadtbaugeschichte, Fakultät 6, Institut für Architektur, TU Berlin

Die historische Bauforschung widmet sich der sachgemässen, quantitativ auswertbaren Aufnahme der äusseren Form der Objekte

- Institut für Kultur und Ästhetik digitaler Medien, Leuphana Universität Lüneburg

Der Projektpartner ist für den Einsatz von HyperImage als zentrales Erfassungs- und Visualisierungsmedium für den Topographie-Visualisierer zuständig.

- DAASI International GmbH, Tübingen

DAASI verantwortet die Schnittstelle zwischen den diversen Datenquellen, bereitet die Daten der Partner auf und konvertiert sie zwecks Import in den Topograp

Wie lautete die Fragestellung zu dem Projekt?

Thomas Kollatz: Im Vorfeld wurden eine Vielzahl von Forschungsfragen zu räumlichen Aspekten formuliert:

Raumproduktion

- Wie verhält sich das Einzelobjekt zum lokalen, räumlichen Ensemble der Einzelobjekte und dem weiteren Umfeld? Lassen sich fachspezifische Erkenntni Raumproduktion vor Ort wiederfinden?
- Sind unterschiedliche Beerdigungsstrategien (Cluster) erkennbar?
- Werden separate Gruppenfelder gebildet z.B. für Wöchnerinnen, Kinder, Unverheiratete, Ehepaare, Opfer von Epidemien (Pest, Cholera), Weltkriegsgefal
- Gibt es visuelle Bezüge zwischen Familien/Ehepartnern bei chronologischer Beerdigung (identische Gestaltung der verwendeten Ornamentik und Symbol derselbe Steinmetz beauftragt)?

- Wie wurde der Friedhof in zeitlicher Perspektive belegt?
- Verhältnisse bei von mehreren Gemeinden genutzten Verbandsfriedhöfen (bis zu 25 bestattende Gemeinden in Franken)?
- Wie ist das Grabmal positioniert? Hat die Position und Ausrichtung religiöse oder repräsentative Gründe (Ausrichtung gegenüber den Himmelsrichtungen bestimmten Punkten etc.)?

Makroebene

- Lassen sich auf Makroebene (überregional, wie Herkunft der Bestatteten, Vergleichsbeispiele anderer Friedhöfe) zu gewissen Zeitpunkten Typologien dur Material (Sandstein, Muschelkalk, dunkles Hartgestein etc.) bzw. Inschrift erkennen, die auf familiäre, soziale, religiöse und topographische Herkunft zuri Sind dadurch Migrationbewegungen der jeweiligen Gesellschaft in den Objektensemblen ablesbar? (Sepharden, Aschkenasen, Herkunft aus Ost- oder Süd bestimmte Familienverbände etc.)? Ist die Form des Grabmales an den sozialen oder religiösen Status gebunden (bestimmte Grabmalformen für Rabbiner ein Zusammenhang zwischen dem Wandel formaler Aspekte des Grabmales aus dem sozialen und religiösen Wandel nachweisen (Veränderung von Grabr Zuwanderergruppen)? Wie funktioniert die Interaktion zwischen jüdischer und nicht-jüdischer Sepulkralarchitektur?

Epigraphik

- Sind idiomatische eulogische Formulierungen an bestimmte Positionen/Felder gebunden?
- Erlaubt die Umgebung eines undatierten Einzelobjektes Hinweise auf dessen zeitliche Einordnung?

Bau- und Kunstgeschichte

- Durch welche Kriterien lassen sich Grabsteintypologien abgrenzen (Bautyp und Aufbau, Form bestimmter Bauteile, Art der Fügung, Materialität etc.)? En Grabsteingestaltung (Grabmaltypologie): Welche Typen dominieren die Grabsteinlandschaft und welche werden nur sporadisch verwendet? Inwiefern hat selbst symbolische Bedeutung? Woher stammen die Vorbilder für bestimmte Grabmaltypen? Ermöglichen bestimmte Materialien spezifische formale Aus
- Industrie- und handwerksgeschichtlich: Woher stammen die Materialien? Welche Steinmetze waren wann, an welchen Grabfeldern und -stellen tätig? Wel Entwicklungen und Regelmäßigkeiten der Anwendungen in Bearbeitungsmethoden (z.B. geätzt, bossiert, usw.) können festgestellt werden? Handelt es sic oder schon Industriebetriebe? Gab es fertige Rohlinge oder handelt es sich um Einzelstücke?

Denkmalpflege

- Welche Natursteinarten wurden wann verwendet und welche Natursteingrabmale sind daher anfälliger als andere? Verursachen bestimmte Grabmaltypen a spezifischen Fügung bestimmte Schadensmuster (mehrteilige und einteilige Grabmaltypen, Art der Verbindungen)? Bei welchen Grabmalen ist welcher Restaurierungsaufwand angemessen (Material, Bautyp, Standsicherheit)? Welche Maßnahmen (z.B. Freilegen der Steine) sind an welchen Grabmalen sch welchen förderlich?
- Restaurierung: Wo besteht akuter Handlungsbedarf (Wichtige Frage für die in der Regel beteiligten Denkmalbehörden und die allgemeine Zugänglichkeit Können Humusschichten, Flechten, Moose zu einem konstanten Milieu beitragen, das längeren Erhalt der Sediment- und metamorphen Gesteine ermöglic abgedeckte und am Boden liegende Grabsteine besser erhalten?
- Welche Informationen und Zusammenhänge sind für die “museumspädagogische” Vermittlung geeignet? Wie müssen die komplexen Informationszusamr Site-Management-System bzw. Besucherinformationssystem aufbereitet und ausgewählt werden?

Interdisziplinär

- Wie funktioniert das Zusammenspiel von Inschriftentext sowie Grabmalform bzw. Anordnung der Symbolik? Besteht ein Zusammenhang zur formalen Ar dem Inhalt der Inschrift? Wie lässt sich ein Zusammenhang quantitativ bzw. qualitativ nachweisen?
- Wurden Symbole zu allen Zeiten durchgängig verwendet? In welchem Verhältnis stehen jüdische, christliche und antike Symbole und Ornamente zueinan Grabsteine von Kohanim (Priesterfamilien) können visuell durch das Symbol der segnenden Hände gekennzeichnet werden. Dies muss aber nicht so sein. die sprachliche – durch Namen oder idiomatische Ausdrücke ("Krone der Priesterschaft" etc.) vorgenommene Differenzierung. Visualisierung des textlich kunstwissenschaftlichen Befundes kann zur Klärung beitragen. Zum Beispiel auch: Ab wann gilt ein sechszackiger Stern als Symbol der Religions-/Ethni (Davidstern) und nicht mehr als Namenssymbol (David) oder reines Schmuckelement?

Warum wurden digitale Methoden gewählt?

Mit analogen Mitteln lässt sich mit erheblichem Aufwand gewiss auch ein statischer Plan erstellen, der allerdings jeweils für genau eine Fragestellung genutzt w die Fülle an raumbezogenen Fragestellungen war allerdings ein digitaler, interaktiver, frei bespielbarer Lageplan erforderlich.

Wie wurden die Daten erhoben?

Thomas Kollatz: Die Daten lagen zum teil schon vor im strukturierten Format EpiDoc: TEI XML für epigraphische Daten. Allerdings war TEI XML für die fein Beschreibung der Kunstwissenschaft und historischen Bauforschung, die sich der äusseren Form der Grabmale unzureichend, so dass im Rahmen des Projektes prototypische XML Auszeichnungmodell für Objektformen entwickelt wurde.

Welche Tools haben Sie ausgewählt und warum? *Thomas Kollatz*: Für den Topographie-Visualiser haben wir die Open-Source Software HyperImage verwen Kunstwissenschaften vielfach Verwendung findet. Im Projekt haben wir diese Software weiterentwickelt und mit Hilfe einer LDAP-Datenbank mit einem Searcl verbunden. Auf diese Weise kann der Lageplan beliebig je nach Fragestellung gefüllt werden.

Wie verlief die Analyse?

Thomas Kollatz: Die meisten der im Vorfeld formulierten Forschungsfragen liessen sich mit dem Topographie-Visualisierer klären. Besonders ertragreich war di Zusammenarbeit zwischen Epigraphik und Bauforschung.

Wie wurden die Ergebnisse publiziert?

Thomas Kollatz: Search-Interface und Topographie-Visualizer werden online zugänglich sein. Sämtliche Projektberichte stehen auf der Projekthomepage. Zuder Veröffentlichung der Projektergebnisse in den DARIAH Working Papers geplant.

Weitere Informationen zum Projekt: https://dev2.dariah.eu/wiki/display/RIRPUB/RiR

Björn Ommer und Peter Bell: Analyse kunsthistorischer Bilddatensätze

Björn Ommer, Professor für Computer Vision der Universität Heidelberg und Direktor des Heidelberg Collaboratory for Image Processing (HCI) und *Peter Bell* Kunsthistoriker und wie Ommer WIN-Kollegiat der Heidelberger Akademie der Wissenschaften, koordinieren seit 2011 gemeinsam bildwissenschaftliche Projel Schnittstelle zwischen Kunstgeschichte und Computer Vision.

Error creating thumbnail: File missing

Passion Search Prototype of an unrestricted image search of the crucifixion: http://hci.iwr.uni-heidelberg.de/COMPVIS/projects/suchpassion/

Wie ist die Idee zu dem Projekt entstanden ?

Peter Bell: Prof. Ommer und Prof. Liselotte Saurma begannen im Rahmen der Exzellenzinitiative eine interdisziplinäre Zusammenarbeit, um die Forschungen b Bildwissenschaften zu verknüpfen.

Wie lautete die Fragestellung zu dem Projekt?

Peter Bell: Mittelalterliche Buchmalerei mit wiederkehrenden Motiven sollte anhand von Algorithmen erschlossen werden. In Folgeprojekten kamen Fragen zu Kommunikation in Bildern (Gesten, symbolische Kommunikation) und grundlegende Überlegungen zur kognitiven und semantischen Erschließung von kunst- u architekturhistorischen Datensätzen auf.

Wie wurde das Projekt finanziert?

Peter Bell: Die Projektgelder kamen aus Mitteln der Exzellenzinitiative (Frontier-Projekte), dem Juniorprofessorenprogramm des MWK BaWü und der Heidelb der Wissenschaften (http://hci.iwr.uni-heidelberg.de//COMPVIS/projects/). Ein DFG-Projekt in Kooperation mit Prometheus ist in Planung. Eng arbeiten wir au Arbeitskreis digitale Kunstgeschichte zusammen. Bei jedem Projekt ist circa ein halbes Jahr Vorlauf gewesen, viele Projekte bauen dabei aufeinander auf (nicht Bereich der Digital Humanities sondern auch von Seiten der informatischen Projekte).

Mit welchen Partnern wurde das Vorhaben umgesetzt und wie waren die Rollen verteilt?

Peter Bell: Im ersten Frontier Projekt war Prof. Saurma Partnerin für die Architektur gewannen wir Prof. Hesse. Seit meinem Einstieg in das Projekt 2011 lag di Organisation der Projekte bei Björn Ommer und mir, unterstützt wurden wir von nahezu allen DoktorandInnen der Gruppe (insb. Takami, Monroy, Arnold) und Postdoc J. Schlecht.

Warum wurden digitale Methoden gewählt?

Peter Bell: Die digitalen Methoden bzw. die Zusammenarbeit mit der Informatik war von Seiten der Kunstgeschichte angeraten, da die wachsenden Bilddatenba textbasiert hinreichend erschlossen werden können. Für die Computer Vision stellten kunsthistorische Datensätze eine interessante und skalierbare Herausforder

Wie wurden die Daten erhoben?

Peter Bell: Die Daten wurden unter anderem von der Universitätsbibliothek Heidelberg zur Verfügung gestellt (Palatina-Handschriften, Sachsenspiegel, Architel Jüngst stellte das Prometheus Bildarchiv (Köln) mehrere tausend Kreuzigungsdarstellungen zur Verfügung, um die Performanz der Algorithmen zu testen.

Welche Tools haben Sie ausgewählt und warum?

Peter Bell: Die Computer Vision Algorithmen wurden in Matlab programmiert und teilweise zu Test- und Forschungszwecken mittels eines PHP-Webinterfaces gemacht. Aufgrund der Komplexität kunsthistorischer Bilddatensätze sollte nicht eine bestehende Lösung adaptiert werden, sondern informatische Grundlagenfo der Datensätze durchgeführt werden.

Wie verlief die Analyse?

Peter Bell: Zur Analyse wurde in fünfzehn verschiedenen Testdatensätze unterschiedliche Objekte und Regionen gesucht und einige ähnliche Bilder in aufwendi im Detail verglichen.

Wie wurden die Ergebnisse publiziert?

Peter Bell: Die Ergebnisse wurden in Zeitschriften und Konferenzbänden publiziert, dabei wurde Wert auf eine Publikation in beiden Fächern gelegt. Die im Prc Bildsuche soll außerdem im Prometheus Bildarchiv genutzt werden.

Weitere Informationen

Heidelberg Collaboratory for Image Processing. Projekte: http://hci.iwr.uni-heidelberg.de/COMPVIS/projects/

Peter Bell, Joseph Schlecht, and Björn Ommer: Nonverbal Communication in Medieval Illustrations Revisited by Computer Vision and Art History, in: Visual R International Journal of Documentation (Special Issue: Digital Art History), 29:1-2, S. 26-37, 2013.

Peter Bell and Björn Ommer: Training Argus, Kunstchronik 68(8): pp. 414-420, August 2015

Antonio Monroy, Peter Bell and Björn Ommer: Morphological analysis for investigating artistic images, in: Image and Vision Computing 32(6), pp. 414-423, 2(

Antonio Monroy, Peter Bell, and Björn Ommer: Shaping Art with Art: Morphological Analysis for Investigating Artistic Reproductions, in: ECCV'12 (VISART) pp. 571-580.

Masato Takami, Peter Bell and Björn Ommer: Offline Learning of Prototypical Negatives for Efficient Online Exemplar, in: Proceedings of the IEEE Winter Co Applications of Computer Vision, pp. 377-384, 2014.

Masato Takami, Peter Bell and Björn Ommer: An Approach to Large Scale Interactive Retrieval of Cultural Heritage, in: Proceedings of the EUROGRAPHICS Graphics and Cultural Heritage, EUROGRAPHICS Association, 2014.

Andrea Rapp: TextGrid - ein Virtueller Forschungsverbund

Andrea Rapp ist Professorin für Germanistische Computerphilologie und Mediävistik an der TU Darmstadt, sie kooperiert in zahlreichen Projekten an der Schni Philologie, Informatik und Forschungsinfrastruktur. 2005 initiierte sie mit 10 weiteren Partnern das Projekt TextGrid - Digitale Forschungsumgebung für die Geisteswissenschaften.

Error creating thumbnail: File missing

Das TextGridLab als digitaler Werkzeugkasten für geisteswissenschaftliche Forschung. Das TextGridLab dient hier als Werkzeugkasten bei der Erstellung einer kommentierten digitalen Edition. Mit dem Text-Bild-Link-Editor können beispielsweise Faksimiles transkribiert und mit Informationen in XML angereichert werden. Die eingescannten Originaltexte können dann mit der Transkription verknüpft werden. http://www.textgrid.de

Wie ist die Idee zu dem Projekt entstanden?

Andrea Rapp: TextGrid ist ein Projekt, das digitale Forschungsinfrastruktur konzipiert, entwickelt und nachhaltig verfügbar hält, insofern ist es nicht ganz vergle klassischen geisteswissenschaftlichen Forschungsprojekten. Einer unserer zentralen Impulse war damals, dass wir Philologie mit digitalen Mitteln betreiben wol nicht die optimalen, standard-basierten OpenSource-Werkzeuge und -Quellen zur Verfügung hatten, die wir brauchten. Hier hat sich natürlich in der letzten Dek getan, dennoch ist dieser Bedarf immer noch groß oder wächst sogar mit steigender Akzeptanz digitaler Verfahren.

Wie lautete die Fragestellung zu dem Projekt?

Andrea Rapp: Die Initiativgruppe bestand in erster Linie aus PhilologInnen und TextwissenschaftlerInnen mit einem starken Fokus auf der Editionsphilologie, ir kollaboratives Arbeiten und digitale Werkzeuge bereits vergleichsweise breit im Einsatz waren, jedoch viele Insellösungen entstanden waren. Unsere Idee war je auf OpenSource und offene Standards zu setzen. In den späteren Phasen wurde neben der Tool-Entwicklung das Community-Building und die Nutzerberatung u immer wichtiger. Als ganz zentral hat sich das TextGrid Repository herausgestellt, das derzeit gemeinsam mit dem DARIAH-DE Repository gepflegt und weiter Inhalte kommen durch die Digitale Bibliothek und durch die Publikationen der NutzerInnen laufend hinein.

Wie wurde das Projekt finanziert? Falls es Förderanträge gab, wie gestaltete sich der Vorlauf?

Andrea Rapp: Die D-Grid-Initiative des Bundesministeriums für Bildung und Forschung (BMBF) bot uns damals die Möglichkeit, unsere Pläne in einem recht u Kontext zu konkretisieren: Als einziges geisteswissenschaftliches Projekt unter natur- und ingenieurwissenschaftlichen Vorhaben haben wir viel von den Forschu dieser Disziplinen und ihrer VertreterInnen profitiert, denn für sie ist der Zugang zu digtaler Infrastruktur etwas Selbstverständliches - vielleicht so wie für uns d der Zugang zu einem Archiv etwas Selbstverständliches sind. Das BMBF bot uns die Chance, die TextGrid-Infrastruktur in drei Förderphasen aufzubauen. Vor a Antrag galt es, sich im ungewohnten Kontext der Infrastrukturentwicklung zurechtzufinden und die Vorgaben der D-Grid-Initiative zu berücksichtigen, was durc Koordination der SUB Göttingen aber gut an alle Partner vermittelt wurde.

Mit welchen Partnern wurde das Vorhaben umgesetzt und wie waren die Rollen verteilt?

Error creating thumbnail: File missing

Das Textgrid-Repository: http://www.textgridrep.de/

Andrea Rapp: TextGrid wurde und wird von einem großen Konsortium getragen. In den drei Förderphasen wechselte jeweils ein wenig die Zusammensetzung. I folgenden Institutionen mit spezifischen Anliegen und Kompetenzen, die sie einbrachten:

- Berlin-Brandenburgische Akademie der Wissenschaften (Nutzerbefragungen und Akzeptanz)
- Technische Universität Berlin (Soziologie, Wissenschaftliche Begleitforschung, Monitoring)
- DAASI International GmbH (Basistechnologie, Authentifizierungsinfrastruktur)
- Technische Universität Darmstadt (TextGrid Laboratory, Schulungen, Beratung)
- Musikwissenschaftliches Seminar der Universität Paderborn/Detmold (Musikwissenschaftliche Tools)
- Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte Duisburg/Essen (Tools)
- Gesellschaft für Wissenschaftliche Datenverarbeitung Göttingen mbH (Basistechnoligie, TextGrid Repository)
- Niedersächsische Staats- und Universitätsbibliothek Göttingen (Verbundkoordination, TextGrid Repository)
- Institut für deutsche Sprache in Mannheim (Linguistische Tools und Korpora)
- Technische Universität Kaiserslautern (OCR-Tool)
- Max-Planck-Digital Library (Tools)
- Max-Planck-Institut für Wissenschaftsgeschichte, Berlin (Tools, Usability, Prozessorganisation)
- Münchener Zentrum für Editionswissenschaft MüZE (Glossen-Editionstool)
- Saphor GmbH (Basistechnologie)
- Universität Trier (TextGrid Laboratory, Wörterbuchnetz)
- Hochschule Worms (TextGrid Laboratory, Tests)
- Julius-Maximilians-Universität Würzburg (TextGrid Laboratory, Digitale Bibliothek)

Die Gruppe hat ferner einen Verein gegründet, um TextGrid weiter gemeinsam betreuen und weiterentwickeln zu können.

Wie wurden die Ergebnisse publiziert?

Andrea Rapp: Das Laboratory als OpenSource Software kann über die TextGrid-Homepage frei heruntergeladen werden. Das Repository bietet eine umfangreic konforme Sammlung zumeist deutschsprachiger Kanonliteratur zur freien Nachnutzung an, die zugleich mittels linguistischer Tools analysierbar ist (hier wird au Suite zurückgegriffen). Zahlreiche Reports, Aufsätze und Bücher finden sich auf der TextGrid-Homepage verzeichnet und zumeist auch als Download verfügbar Informationen gibt es im von TextGrid mitbetriebenen DHd-Blog, in verschiedenen Mailinglisten, im YouTube-DHd-Kanal oder über Twitter @TextGrid.

Weitere Informationen zu den Projekten

http://www.textgrid.de

http://www.textgridrep.de

http://www.dhd-blog.org

https://www.youtube.com/user/dhdkanal

Patrick Sahle: Altägyptisches Totenbuch - ein digitales Textzeugenarchiv

Patrick Sahle ist Geschäftsführer des Cologne Center for eHumanities (CCeH). Dort betreut er im Rahmen einer Kooperation zwischen der Universität zu Köln Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste (AWK) auch die Umstellung schon länger laufender Akademievorhaben auf digitale Art

Publikationsformen. Am CCeH ist in Zusammenarbeit mit der Abteilung für Ägyptologie an der Universität Bonn in den Jahren 2011-2012 mit dem "Altägyptis ein digitales Textzeugenarchiv" die digitale Präsentation des damit zuende gehenden Langzeitvorhabens erarbeitet worden.

Error creating thumbnail: File missing

Das altägyptische Totenbuch: Ein digitales Textzeugenarchiv
http://totenbuch.awk.nrw.de/

Wie ist die Idee zu dem Projekt entstanden?

Patrick Sahle: Die Bonner Arbeiten zum Totenbuch (seit 1994) umfassten schon früh eine Datenbank zur Objektbeschreibung sowie ein Bildarchiv. In den späte wurde deutlich, dass eine "Publikation" der gesamten Projektergebnisse am besten online erfolgen sollte. Das digitale Abschlussprojekt ergab sich damit fast zw dem inhaltlichen Forschungs- und Erschließungsprojekt.

Wie lautete die Fragestellung zu dem Projekt?

Patrick Sahle: Das Akademievorhaben selbst sammelte alle relevanten Informationen zu allen bekannten Überlieferungsträgern zum altägyptischen Totenbuch (von Sprüchen) und leistete hier vor allem Erschließungsarbeit, z.B. in der Identifikation der einzelnen Texte auf den Objekten. Im Digitalisierungsprojekt ging e wie man die komplexen Informationen und das Bildarchiv auf die beste Weise dem forschenden Fachpublikum und einer breiteren interessierten Öffentlichkeit und damit die weitere Beschäftigung mit dem Material fördern könnte.

Wie wurde das Projekt finanziert?

Patrick Sahle: Im Rahmen einer Digitalisierungsinitiative hat die Union der Akademien der Wissenschaften in Deutschland den einzelnen Akademien zusätzlich Umstellung schon länger laufender Vorhaben auf digitale Arbeits- und Publikationsformen zur Verfügung gestellt. Diese Mittel konnten für das Projekt genutzt ein entsprechender Projektplan entwickelt worden war.

Mit welchen Partnern wurde das Vorhaben umgesetzt und wie waren die Rollen verteilt?

Patrick Sahle: Das Vorhaben war eine ganz typische Zusammenarbeit zwischen einer fachwissenschaftlichen Seite und einem DH-Kompetenzzentrum. Für die fachwissenschaftliche Seite der Abteilung für Ägyptologie der Universität Bonn hat vor allem Marcus Müller, aber auch etliche weitere Projektmitarbeiter an de Projekts, an der Diskussion zum Datenmodell und zum Portal teilgenommen. Neben der inhaltlichen Forschungsarbeit haben die Fachwissenschaftler aber auch Bilddigitalisierung und die Dateneingabe selbstständig durchgeführt. Auf der DH-Seite haben sich am CCeH Patrick Sahle, Ulrike Henny, Jonathan Blumtritt un um die Modellentwicklung, die Datenpflege (Übernahme, Konversion, Bereinigung, Anreicherung), die Systementwicklung (Server, Datenbank, Abfrageskripte Nutzerverwaltung, Oberfläche) und die Visualisierungen gekümmert. Um zu einer guten Publikation zu kommen, war außerdem ein Designbüro für den Portale

Warum wurden digitale Methoden gewählt?

Patrick Sahle: Eine Publikation des gesammelten Wissens und der Abbildungen in Buchform wäre offensichtlich weder vom Umfang noch von der Funktionalit gewesen. Nur die digitale Publikation erlaubte vielfältige Browse- und Suchzugänge, zusätzliche Visualisierungen von Zusammenhängen und eine permanente des Wissens.

Wie wurden die Daten erhoben?

Patrick Sahle: Die Daten wurden von den ägyptologischen FachwissenschaftlerInnen zunächst in einer Filemaker-Datenbank erhoben. Das Modell wurde im Digitalisierungsprojekt in XML überarbeitet und verfeinert. Die Altdaten wurden dann migriert, homogenisiert, überarbeitet und von den FachwissenschaftlerIn gepflegt. Zusätzlich wurden nebenläufige "Wissensbasen" angelegt, z.B. zu Materialgruppen, Datierungen, Geographica etc.

Welche Tools haben Sie ausgewählt und warum?

Patrick Sahle: Die gesamte Datenhaltung beruht auf einem lokalen XML-Dialekt. Man hätte lieber einen Standard verwendet, allerdings lag kein passender vor, ein projektinternes Modell und die Zeit drängte. Die Daten wurden (und werden) in einer XML-Datenbank (eXist) verwaltet. Die Datenpflege erfolgte direkt im Mode, der mit der Datenbank verbunden war. Die bereits vorhandenen Abbildungen der Stücke wurde mit einfachen Scannern digitalisiert weil eine Faksimilier eigentlichen Zielen des Projektes gehörte. Weitere Abbildnugen wurden mit digitalen Kameras vor Ort angefertigt.

Wie verlief die Analyse?

*Patrick Sahle:*Das bestehende Filemaker-Modell wurde übernommen, kritisch gesichtet und in der Diskussion zwischen den Fachwissenschaftlern und den DH- Basis von XML weiter entwickelt.

Wie wurden die Ergebnisse publiziert?

Patrick Sahle: Das Ergebnis ist ein digitales Portal. Die XML-Datenbank wird mittels xQuery abgefragt und liefert on the fly generierte Webseiten zurück. Aus Gründen gibt es ein Rollenmodell, das bestimmte Abbildungen nur einem registrierten Nutzerkreis zugänglich macht.

Was sind erwähnenswerte Besonderheiten im Projekt?

Patrick Sahle: Die xQuery-Abfragen der Datenbank wurden auch genutzt, um durch Visualisierungen einen besseren Überblick über die komplexen Daten zu sc Besonderer Wert wurde auf die Nachhaltigkeit des Projekt gelegt. Eine der Grundlagen dafür ist eine enge Verbindung zur weiteren Community der Altertumsk werden z.B. die einigermaßen kanonischen "Trismegistos"-Nummern als Identifikatoren der Objekte und zur Herstellung permanenter Links verwendet oder Di Pelagios/Pleiades eingebunden. Das Projekt ist auch ein Testfall dafür, wie zuende gegangene Forschungsprojekte weiterleben können: Beim Totenbuch besteht Eingabemaske weiter, mit der auch Jahre nach Projektende die Daten korrigiert und weiter gepflegt werden können - was tatsächlich manchmal geschieht.

Weitere Informationen zu den Projekten

Startseite: http://totenbuch.awk.nrw.de/

Beispiel für ein Objekt, mit Permalink: http://totenbuch.awk.nrw.de/objekt/tm57143

Beispiel für Informationsvisualisierung: http://totenbuch.awk.nrw.de/uebersicht/sprueche

Hannah Busch: Handschriften analysieren mit eCodicology

Error creating thumbnail: File missing

Exemplarische Layoutmerkmale einer mittelalterlichen Handschriftenseite. Quelle: Projekt eCodicology.. http://www.ecodicology.org/index.php?id=4

Hannah Busch ist wissenschaftliche Mitarbeiterin am Trier Center for Digital Humanities. Dort arbeitet sie im Verbundprojekt eCodicology, das vom BMBF 20 und gemeinsam mit der Technischen Universität Darmstadt (Projektkoordinator) und dem Karlsruher Institut für Technologie durchgeführt wird.

Wie ist die Idee zu dem Projekt entstanden?

Hannah Busch: In den letzten 10 Jahren wurden viele mittelalterliche Handschriftenbestände ins digitale Medium überführt und einer breiten Öffentlichkeit übe Web zugänglich gemacht. Das Projekt eCodicology entstand aus der Überlegung hinaus, welcher weitere Nutzen – neben der virtuellen Rekonstruktion mittelalt Bibliotheken, deren inhaltlicher Aufbereitung und Präsentation – aus diesen Daten gezogen werden kann.

Error creating thumbnail: File missing

Visualisierung der Layoutmerkmale eines Teils des St.Mattheiser Bestandes

Wie lautete die Fragestellung zu dem Projekt?

Hannah Busch: Ziel von eCodicology ist die Entwicklung, Erprobung und Verbesserung von neuen Algorithmen, die makro- und mikrostrukturelle Gestaltungsr mittelalterlichen Handschriftenseiten erkennen, um deren Metadaten anzureichern. Die Beschreibungen aus früheren Handschriftenkatalogen können auf diese V automatisiert ergänzt und erweitert werden, beispielsweise waren bisher häufig nur besonders nennenswerte Miniaturen lokalisiert, erfasst und beschrieben.

Zusätzlich können mit Hilfe der Bildung von Korrelationen zwischen den Metadaten aus den Handschriftenbeschreibungen (z.B. Datierung, Textgattung, Besch automatisch erhobenen Layoutdaten weitere Fragen zur Entdeckung verborgener Zusammenhänge zwischen Handschriften an den Bestand von St. Matthias ger den vielversprechendsten Ansätzen zählen:

- Die Entdeckung von Abhängigkeiten zwischen beschrifteten/freien Flächen und dem Format der Handschrift
- Die Ermittlung des Verhältnisses von Bildraum und Textraum auf den Seiten
- Das Aufspüren von Bezügen zwischen Textinhalt und Gestaltung der Seiten
- Das Zusammenführen von Fragmenten anhand der Seitengestaltung
- Die Identifikation von Schreiberhänden anhand der Seitengestaltung

Warum wurden digitale Methoden gewählt?

Hannah Busch: Das Festhalten von äußeren Merkmalen auf jeder einzelnen Kodexseite aus einem geschlossenen Handschriftenbestand kann für die Beantwortu Forschungsfragen sinnvoll sein, für die eine sehr große Datenmenge systematisch ausgewertet werden muss. Mit einer rein händischen Vorgehensweise könnten Entwicklungen im Layout des St. Mattheiser Bestandes nur unter großen Anstrengungen verfolgt werden. Eine abschließende statistische Auswertung erlaubt di Handschriftenseiten auf einer empirischen Basis. Regelmäßigkeiten (Muster) bzw. Veränderungen der Layoutkonstellationen lassen sich auf diese Weise aufspü

Im Projekt eCodicology kommt eine Forschungsmethode zum Einsatz, die mit dem in der Computerphilologie etablierten Verfahren des distant reading vergleic detaillierte und intensive close reading einzelner Bücher steht im Vordergrund. Stattdessen soll eine quantitative Gesamtschau über alle Seiten des mittelalterlich der Abtei St. Matthias erfolgen, und zwar hinsichtlich der formalen Gestaltung der Handschriftenseiten. Eine solche Analyse erlaubt es dem Kodikologen, das M Vogelperspektive zu betrachten. Der subjektive Blick des Handschriftenforschers kann auf diese Weise objektiviert werden.

Wie wurden die Daten erhoben?

Hannah Busch: Datengrundlage bilden die im Projekt "Virtuelles Skripotorium St. Matthias" erstellten Digitalisate von rund 440 Handschriften der mittelalterlic der Abtei St. Matthias in Trier. Durch Verfahren der Bildbearbeitung werden die 170.000 Handschriftenseiten zunächst vorprozessiert, um eine Vergleichbarkeit Für die automatisierte Extraktion und Vermessung der Layoutmerkmale in den Digitalisaten werden spezielle Algorithmen angepasst und bei Bedarf entwickelt. Untersuchung wurden einfache Parameter wie Seitengröße, Schriftraum, Bildraum, Abstände und Ränder (freigelassener Raum), graphische Elemente, Anzahl d sowie Anzahl der Textzeilen identifiziert, für die v.a. ihre Ausdehnung (Höhe x Breite) sowie Anzahl und Position (Koordinaten) auf jeder Seite verlässlich gem können.

Welche Tools haben Sie ausgewählt und warum?

Hannah Busch: Die Layoutvermessung wird mit ImageJ durchgeführt, Plugins erlauben eine einfache Anpassung des Workflows an neue Anforderungen und Ar
Zur Speicherung der Ergebnisse wird TEI P5 konformes XML verwendet, das im Bereich der Handschriftendigitalisierung bereits etabliert ist und daher eine ho
mit den Metadaten anderer Digitalisierungsprojekte verspricht. D3 (Data Driven Documents) bietet umfangreiche Visualisierungsfunktionalitäten, die vielfältig
ebenfalls leicht adaptierbar sind.

Wie verläuft die Analyse?

Hannah Busch: Die bei der Bildverarbeitung gewonnenen Daten zu den Layoutmerkmalen werden als sogenannte XML-Tags automatisch in den Metadaten abg
festgehaltenen Werte lassen sich anschließend statistisch auswerten, visualisieren, zueinander in Beziehung setzen und dienen als Grundlage für neue wissenscha
Erkenntnisse. Während des gesamten Prozesses findet zur Verbesserung des Workflows und Beantwortung der Forschungsfragen ein reger Austausch zwischen
statt.

Das Projekt hat einen Modellcharakter, der die Nachnutzung der Ergebnisse anstrebt. Die an den Beständen von St. Matthias erprobten Algorithmen können so a
für die Untersuchung weiterer Handschriftendschriftenbestände dienen.

Wie wurden die Ergebnisse publiziert?

Hannah Busch: Die Handschriftendigitalisate sind bereits über das Projekt „Virtuelles Skriptorium St. Matthias“ im Web frei zugänglich und auf Basis der bishe
Katalogdaten durchsuchbar. In einem weiteren Schritt wurden die Bilddateien mit den dazugehörigen XML Dateien in das TextGrid-Repositorium eingespeist u
beispielsweise zur Transkription oder Bildannotation zur Verfügung. Die Ergebnisse der Recherche und statistischen Auswertung sowie Möglichkeiten zur dyna
Visualisierung werden über das Projektportal bereitgestellt.

Weitere Links:

http://www.ecodicology.org www.stmatthias.uni-trier.de

https://textgrid.de/

DH-Projekte in Europa

Auf Europäischer Ebene gibt es verschiedene Projekte, welche für die Geisteswissenschaften viele Anreize und Unterstützung bieten.

Hier werden virtuelle Forschungsumgebungen für verschiedene geisteswissenschaftliche Domains erstellt, Best Practice Netzwerke gebildet oder Tools und Die
geisteswissenschaftliches Arbeiten entwickelt. Bei EU-weiten Projekten kann man zwischen generischen Projekten und Infrastrukturprojekten unterscheiden. Be
dass sie durch Konsortien aus Partnern von mindestens 3 Ländern betrieben werden.

Die Projekte verfolgen unterschiedliche Ziele und sind im ganzen Spektrum der geisteswissenschaftlichen Forschung anzutreffen. So geht es zum einem um die
Aufbereitung und den Zugang von analogen Quellen als Grundlage für die geisteswissenschaftliche Forschung (z.B. Europeana Research[1]) oder um die Entwi
digitaler Werkzeuge für die Beantwortung von Forschungsfragen und um die Formulierung von Policies und Strategien um die Nachhaltigkeit von geisteswissen
Forschungsinfrastrukturen und den darin vorgehaltenen Forschungsdaten zu befördern (z.B. Parthenos [2]).

Weitere Beispiele für Projekte sind:

- ENARC: The European Network on Archival Cooperation ist ein von der ICARUS4all (International Centre for Archival Research for All) Initiative betri
  Projekt mit Partnern aus hochrangigen Archiven mehrerer europäischer Staaten. Es werden sowohl Digitalisierungsvorhaben durchgeführt als auch Tools
  mithilfe derer die resultierenden Daten angereichert und präsentiert werden können. [3]
- Openaire: Open Access Infrastructure for Research in Europe entwickelt ein übergreifendes Suchportal zum Recherchieren und Zugreifen auf diverse Dat
  Projektkontext. Hier findet der/die geneigte Forschende Daten aus Abschlussarbeiten, Forschungsprojekten, Projektmetadaten und Personen. [4]
- DiXiT: Digital Scholarly Editions Initial Training Network ist ein Bildungs- und Forschungsprogramm für europäische Doktoranden und Postdoktorander
  von digitalen Editionen arbeiten.[5]

Während Projekte tendenziell nach einem festen Zeitraum abgeschlossen sind, streben Infrastrukturen eher eine dauerhafte Förderung -beispielsweise im Rahme
europäischen Organisationsform European Research Infrastructure Consortium (ERIC)- an, damit sie auch unbefristet ihre Dienstleistungen zur Verfügung stelle
Recherche über angebotene Dienste und Portale lohnt sich bei einem eigenen Projektentwurf allemal – gerade auch um potentielle Partner oder unterstützende P
Allerdings ist bei abgeschlossenen Projekten mitunter Vorsicht geboten, da hier der Fall auftreten kann, dass Angebote nicht mehr eingehalten werden und Softw
gepflegt und weiter entwickelt wird.[6]

Eine beispielhafte Übersicht über europäische Infrastrukturprojekte und ihre Förderung bietet die folgende Liste:

EHRI

"The European Holocaust Research Infrastructure"[7] ist ein Verbund aus 20 Institutionen, die aus den Bereichen Holocaustforschung, Archivwissenschaft und
Humanities kommen. Ziel des Projekts ist die digitale Zusammenführung verteilter archivalischer Ressourcen in einem Portal, das WissenschaftlerInnen und Int
Zugang zu den Materialien der Holocaustforschung bietet, diese mit Daten anreichert und auffindbar macht, damit sie mit digitalen Werkzeugen bearbeitet werd

ARIADNE

„Advanced Research Infrastructure for Archaeological Datasets Networking in Europe"[8] verfolgt das Ziel, vorhandene Forschungsdatenbanken in der Archäol
zusammenzuführen, um digitale Daten aus den Bereichen Archäologie und Kulturerbe auf europäischer Ebene wissenschaftlich nutzen zu können.

Das EU-Projekt bringt 24 Partner aus 13 Ländern zusammen. Diese teilen archäologische Datenbanken und Kenntnisse im Bereich der Datentechnik, um die Ba
europäische Forschungsinfrastruktur zu schaffen. Sie soll zukünftig der archäologischen Wissenschaft wie auch dem Kulturgutmanagement von Nutzen sein. Zi
sind, verwendete Daten- und Metadatenstandards zusammenzustellen, gemeinsame Mindeststandards zu erarbeiten und Schnittstellen zu entwickeln.

ESFRI

steht für the European Strategy Forum on Research Infrastructures und ist ein Förderbereich der EU für nachhaltigen Zugang zu virtuellen Forschungsinfratsruk
zählen fünf pan-europäische Infrastrukturen in den Sozial -und Geisteswissenschaften (SSH - Social Science and Humanities) zu den ESFRI-Infratstrukturen[9]:
ERIC[10], CLARIN-ERIC[11], ESS-ERIC[12], CESSDA[13], SHARE-ERIC[14]. Hier werden beispielhaft nur DARIAH und CLARIN vorgestellt.

DARIAH-ERIC

Das Digital Research Infrastructure for the Arts and Humanities (DARIAH) ERIC möchte Wissenschaftler, Werkzeuge und digitale Methoden in den Digital Hu zusammenbringen, interdisziplinäres Arbeiten in den Geisteswissenschaften unterstützen und nachhaltige Strategien für Forschungsdaten entwickeln.

CLARIN-ERIC

Common Language Resources and Technology Infrastructure (CLARIN) ist mit Ausnahme von DARIAH das einzige andere geisteswissenschaftliche Forschungsinfrastrukturvorhaben, welches im Rahmen der ESFRI-Roadmap von der EU bewilligt wurde. Die Nutzer der CLARIN-Infrastruktur stammen aus d Sozialwissenschaften und beschäftigen sich mit sprachbasierten Ausgangsdaten. Abhängig von fachlichen Vorlieben und Netzwerken kann für Forschende im B Humanities CLARIN die benötigten Infrastrukturen bereitstellen. Zugrundeliegende Kernfunktionen werden häufig von CLARIN und DARIAH gemeinsam ger und vorangetrieben.

Anmerkungen

1. ↑ Europeana Research möchte WissenschaftlerInnen den Zugang zu in der Europeana referenzierten Materialien erleichtern, damit diese als Grundlage für For werden können. Dabei geht es um Klärung von rechtlichen Fragen und Zugangsmodalitäten. http://research.europeana.eu/
2. ↑ http://www.parthenos-project.eu/
3. ↑ http://enarc.icar-us.eu/
4. ↑ Vgl. www.openaire.eu
5. ↑ http://dixit.uni-koeln.de/
6. ↑ Für Übersichten vgl. http://humanum.hypotheses.org/155, https://dariah.eu/about/collaboration.html, http://de.slideshare.net/dri_ireland/peter-doorn
7. ↑ http://www.ehri-project.eu/
8. ↑ http://www.ariadne-infrastructure.eu/
9. ↑ https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri
10. ↑ https://dariah.eu/
11. ↑ http://clarin.eu/
12. ↑ http://www.europeansocialsurvey.org/
13. ↑ http://cessda.net/
14. ↑ http://www.share-project.org/

Vom Datenberg zum Wissensfluss: Wie organisiert man Forschungsdaten?

Die Digital Humanities zeichnen sich dadurch aus, dass sie digitale Daten generieren und/oder den Erkenntnisprozess auf Daten dieser Art aufbauen. Grundlage können analoge Inhalte, wie beispielsweise Quellen, Manuskripte, Gemälde, etc. sein, die digitalisiert werden oder mit digitalen Methoden untersucht werden. D entstehenden Daten sind vielfältig in ihren Formaten, Funktionen und repräsentierten Inhalten, was eine weitere Spezifizierung sinnvoll macht.

Grundsätzliches zuerst: Zur Definition von Daten und ihrem Entstehungskontext

Digitale Daten lassen sich nach verschiedenen Gesichtspunkten und Perspektiven unterscheiden. Dabei spielt (a) der Kontext der Entstehung der Daten, (b) ihre innerhalb des geisteswissenschaftlichen Forschungsprozesses sowie (c) die inhaltlich Ausrichtung der Daten eine ausschlaggebende Rolle. Aus dieser Untersche unterschiedliche Anforderungen an die Daten und deren AnbieterInnen ableiten.

(a) Kontext: Entstehung und Provenienz der Daten

Zum einen können digitale Daten entstehen, wenn analoge Inhalte digitalisiert werden oder wenn Daten digital erstellt werden beispielsweise mittels Software u Aufnahmegeräten. Bei der Digitalisierung physischer Objekte oder Inhalte kommt es vor allem auf die Genauigkeit des Digitalisats an und inwieweit es als Stell Surrogate für die analogen Inhalte fungieren kann. Hier ist es wichtig zu unterscheiden, inwieweit die digitale Rekonstruktion Eigenschaften des physischen Obj abbilden kann. Die Tiefe und Genauigkeit der Digitalisierung bestimmt auch, welche Forschungsfragen damit beantwortet werden können.[1] So liefert die Digit Buches für einen/eine TextwissenschaftlerIn meist keinen Informationsverlust, der/ die GeisteswissenschaftlerIn dessen Forschungsobjekt aber das Buch und sei sich ist, wird kaum sein/ihr Forschungsinteresse nur auf die digitale Version des Buches beschränken wollen. Ein weiterer wichtiger Punkt sind die digital erzeu auch born-digital[2]), die innerhalb des Forschungsprozesses entstehen oder als Basis für den Erkenntnisgewinn dienen. Einerseits können die z.B. Social-Media anderseits sind dies Daten, die mit digitalen Werkzeugen erstellt wurden, z.B. Bilder, Annotationen, Fragebögen, GIS-Daten aus Geoinformationssystemen oder Office-Anwendungen[3].

(b) Funktion innerhalb des geisteswissenschaftlichen Forschungsprozesses

Im Hinblick auf Daten als Grundlage für geisteswissenschaftliche Forschung und Ausgangspunkt des Erkenntnisprozesses wird oft von Forschungsdaten gespro zu definieren, ist schwierig und häufig kommt es auf die Funktion der Daten innerhalb des Forschungsprozesses an. Dies zeigt sich besonders in der Abgrenzung und Sekundärdaten. Puhl et al. [4] sagen, dass die Grenzen zwischen Primär- und Sekundärdaten fließend verlaufen und nur aus der Perspektive des jeweiligen Forschungsprozesses heraus bestimmt werden können (S. 9). Daraus ergibt sich, dass Primärdaten herangezogen werden, um gegebene Forschungsfragen zu bea während Sekundärdaten schon als Ergebnis eines Erkenntnisprozesses vorliegen (ebd, S. 9). Wir folgen hier der Definition von Puhl at al.:

> „Unter digitalen geistes- und kulturwissenschaftlichen Forschungsdaten werden innerhalb von DARIAH-DE all jene Daten verstanden, die im Kontext ein und kulturwissenschaftlichen Forschungsfrage gesammelt, beschrieben, ausgewertet und/oder erzeugt wurden."
> – Puhl, Andorfer, Höckendorff, Schmunk, Stiller und Thoden: Diskussion und Definition eines Research Data Life Cycle für die digitalen Geisteswisse

Andorfer[5] zeigt im Zuge der Auswertung von Interviews mit FachwissenschaftlerInnen, dass der Begriff der Forschungsdaten wenig Verwendung findet, wenn oder Publikationen gemeint sein. Im Hinblick auf die Nachnutzbarkeit von Daten, die während des Forschungsprozesses erstellt werden, waren die interviewten WissenschaftlerInnen jedoch sehr wohl bereit ihre Daten zu publizieren und anderen zur Verfügung zu erstellen, als auch selbst auf solche "Forschungsdaten" zu

(c) Inhaltliche Ausrichtung der Daten

Daten lassen sich auch noch hinsichtlich ihrer inhaltlichen Ausrichtung unterscheiden: handelt es sich um Daten oder Metadaten. Metadaten sind ganz generell I die die technischen, administrativen oder funktionellen Eigenschaften dieser beschreiben. Diese können selbst Forschungsdaten sein, wenn beispielsweise biblio Angaben für bibliometrische Analysen genutzt werden. Wichtig sind dabei auch Nachweisinstrumente für Forschungsdaten, auch wenn diese nur analog vorlieg aber digital sind.

Wenn Daten einen Sinnzusammenhang bilden, spricht man auch von digitalen Objekten, die sich durch eine gewisse inhaltliche Zusammengehörigkeit auszeichn Objekt beispielsweise kann ein Digitalisat mit seinen entsprechenden Metadaten und angereicherten Vokabularen sein. Alle Daten innerhalb eines digitalen Obje unterschiedlichen Formaten mit unterschiedlichen Lizenzen vorliegen.

Um Handlungsvorgaben und Empfehlungen für den Umgang mit Daten in Forschungsumgebungen zu liefern, wurden viele Referenzmodelle für den digitalen Forschungsdatenkreislauf entwickelt. DARIAH-DE hat solch einen Datenzyklus aufgestellt, um die verwendeten Daten, ihre Bearbeitung innerhalb geisteswisse Forschungsaktivitäten und daraus resultierende Ergebnisse transparenter zu machen und Handlungsempfehlungen für die DARIAH-Infrastruktur zu liefern[6]. A Langzeitarchivierung, Publikation und Nachnutzung von Daten fanden auch berücksichtig und sind in den Kreislauf eingeflossen.

Error creating thumbnail: File missing

Die Abbildung zeigt den entwickelten Research Data Life Cycle, die verschiedenen Arbeitsschritte, die daraus resultierenden Datenformate und die Voraussetzungen für eine Nachnutzung und Langzeitarchivierung der Daten.

Aus Masse mach Klasse - aber wie? Interoperabilität durch Standardisierung

In einigen geisteswissenschaftlichen Fachdisziplinen haben sich in den vergangenen Jahren fachspezifische nationale und international angewandte Standards he denen bestimmte Kategorien von Forschungsdaten, Forschungsobjekten oder auch Forschungsprozessen beschrieben werden können. Festzustellen ist auch, das Vielzahl von Editionsprojekten – unabhängig von ihrer disziplinären Verortung – solche Standardisierungsprozesse herausgebildet haben. So sind hier an erster S *Encoding Initiative* (TEI) mit einem internationalen Spektrum oder auch das deutsche Projekt TextGrid zu nennen, die maßgeblich zu einheitlicheren Verfahren

Als weiterer Aspekt ist zu nennen, dass in der Vergangenheit von Seiten der Drittmittelgeber für ausschließlich inhaltlich orientierte und ausgerichtete Forschung Erfassung und Erschließung des verwendeten Quellenmaterials mit standardisierten Metadaten nicht explizit gefördert wurden. Aber auch hier sind in den letzte beispielsweise durch das Förderprogramm für wissenschaftliche *Literaturversorgungs- und Informationssysteme* (LIS) der Deutschen Forschungsgemeinschaft I entwickelt worden, die maßgeblich zu Standardisierungen und einer interdisziplinären Interoperabilität beigetragen haben. Trotz dieser Einschränkungen ist erke bei den zuletzt genannten Disziplinen eine stärkere Verwendung von Normdaten, wie beispielsweise des *Thesaurus of Geographic Names* (TGN) oder der *Geme Normdatei* (GND), erkennbar ist und diese Entwicklungen auch auf Dauer unterstützt werden sollten.

Zur Notwendigkeit semantischer Auszeichnungen

Eine Maschine dürfte Schwierigkeiten haben, im folgenden Satz „Hans Hamburger genießt in Paris einen Berliner“ Ortsnamen von Personennamen und populär unterscheiden. Beißt Hans Hamburger in Paris/Frankreich oder in Paris/Texas in die Süßspeise? Hier könnte eine Spezifizierung über den Getty Thesaurus of Ge für geographische Präzisierung sorgen. Gleiches gilt für Hans Hamburger, denn immerhin gibt es drei Personen dieses Namens mit je eigener *Identifikator* (ID) *Gemeinsamen Normdatei* (GND) der deutschen Nationalbibliothek und diese gilt es dann zu spezifizieren. Und um auszuschließen, dass mit „Berliner“ Kennedy Berliner“ assoziiert wird, könnte durch Hinweis auf Kategorie 642 „Mahlzeiten und Tischkultur“ der Deweyschen Decimal Classification auf die Backware hing

Eine einfache Suche bei Wikipedia kann ebenfalls die Unklarheiten bzw. Probleme veranschaulichen, die beim Text-Mining und dem Matching auf bestimmte V entstehen – zum Beispiel bei der Erkennung von Homonymen; siehe dazu etwa die Begriffsklärung zu „London“ in der deutschsprachigen Wikipedia.[7]

Datenqualität

Die Qualität der Daten bestimmt maßgeblich, inwiefern diese für die Forschung nutzbar und nachnutzbar sind.

So kann die Auflösung von Bilddaten für die Beantwortung einiger Forschungsfragen sehr entscheidend sein, während für die Beantwortung anderer Forschungs beispielsweise eher die Qualität der Metadaten, bspw. im Bitstream eines Bildes, bedeutend sein kann.

Auf der einen Seite spielt die Datenqualität eine Rolle für das Auffinden von digitalen Objekten auf der anderen Seite ist sie essentiell für die Analyse der Daten Methoden. Die Qualität von Metadaten wird oft von Nutzern und Anbietern von Metadaten unterschiedlich betrachtet und eingeordnet. So beschreibt zum Bespi Europeana[8] gute Metadatenqualität als Voraussetzung, um digitalisierte kulturelle Objekte zu beschreiben, zu finden und überhaupt mit Ihnen weiter arbeiten z bibliothekarischen Bereich wurden auch einige Anstrengungen unternommen, um den Begriff der Metadatenqualität näher zu beschreiben und Handlungsempfe Verbesserung von Metadatenqualität erstellen zu können.[10]

Bereits bei der Digitalisierung und Aufbereitung ist auf die Qualität der Daten zu achten. Hier ist beispielsweise die Fehlerrate der Texterkennung (*Optical Char* oder OCR) in digitalisierten Volltexten zu berücksichtigen - Textmining-Ergebnisse basierend auf digitalisierten Volltexten nicht zu hinterfragen und die Qualiät kennen kann gefährlich sein. [11]

Weiterhin ist Datenqualität essentiell für die Nachnutzung der Forschungsdaten seien dies nun Metadaten oder die Daten selbst. Hier ist es auch selbstverständlic Richtlinien guter wissenschaftlicher Praxis eingehalten werden und sich diese auch in den publizierten Daten widerspiegeln.[12] Eine Sicherung der Daten in ins Repositorien, die fachspezifische Richtlinien umsetzen und eine Bereithaltung der Daten garantieren, unterstützt auch die Einhaltung von Qualitätsstandards. Ki unterscheidet neben der Qualität der Daten und Metadaten auch noch eine dritte Ebene, nämlich die Qualität der Forschungsdateninfrastrukturen. Forschungspro *Registry of Research Data Repositories* (r3data)[14] versuchen hier WissenschaftlerInnen Anhaltspunkte für die Einschätzung der Güte und Qualität von Reposit

Kontrollierte Vokabulare

Viel Aufmerksamkeit richtet sich bei der Arbeit mit geisteswissenschaftlichen Daten auf die Kategorisierung und Schematisierung der Inhalte. Dies ist gerade au Vielfalt und Verschiedenheit der Daten geboten und es lässt sich wertvolle Arbeitszeit sparen, wenn auf bereits stattgefundene Arbeiten zurück gegriffen werden

So existieren bereits für die Einteilung von personen- und ereignisbezogenen Inhalten so genannte Normdaten oder kontrollierte Vokabulare, mithilfe derer eine von Daten erfolgen kann.

Die Potentiale von kontrollierten Vokabularen sind erheblich: Durch die Verwendung von Semantic Web Strategien können z.B. in der prosopographischen Fors Bezeichnungen für eindeutige historische Personen erkannt und aufgelöst werden und so umfassendere Nachweise erstellt und verwendet werden. Ähnliche Stra bereits für Verwandtschaftsbeziehungen, biographische Informationen (Lebensdaten, Wirkungsorte, Berufe, soziale Rollen) aber auch eindeutige Ortsbezeichnu u.ä. angewandt.

Insbesondere in Kombination mit personenbezogenen Normdaten können so komplexe Zusammenhänge und Vergleichsperspektiven erschlossen werden: Korre Statusgruppe, Konfession, politischer Funktion können ebenso erfasst werden wie die Verortung Einzelner in Personen-, Berufs-, Patronage- und Familiennetzw

Was sind kontrollierte Vokabulare? – Einige Grundzüge

Kontrollierte Vokabulare sind Sammlungen von Wörtern und Bezeichnungen, die nach festgelegten Regeln bearbeitet wurden, um die Mehrdeutigkeiten der n Sprache zu reduzieren. Kontrolliert-strukturierte Vokabulare können zur terminologischen und zur begrifflichen Kontrolle in der Informationspraxis zu Indexier genutzt werden. Als terminologische Kontrolle wird dabei die Möglichkeit genutzt, durch Sammlung von Wörtern, die nach festgelegten Regeln bearbeitet wurd Mehrdeutigkeiten der natürlichen Sprache zu reduzieren. Als begriffliche Kontrolle kann der Aufbau von Relationen zwischen Begriffen verstanden werden. Ko Vokabulare ermöglichen dabei die inhaltliche Erschließung von Dokumenten durch

- eine konsistente Indexierung von gleichartigen Bestandteilen,
- verbesserte Wiederauffindbarkeit von Arten/Datafakten,
- Hilfe bei der Präzisierung der Recherche,
- Verständigung über die Inhalte einer (Wissens-) Domäne,
- Unterstützung der Interoperabilität von Datafakten/Artefakten und
- besseres Verständnis der Semantik von Daten.

Der Zweck von kontrolliert-strukturierten Vokabularen liegt also maßgeblich in der **Organisation von Daten, bzw. darin enthaltenen Informationen**. Kontrol können nach Art und Grad ihrer Strukturierung typologisiert werden. Man kann unterscheiden zwischen:

- einer einfachen Form ohne begriffliche Strukturierung (z.B. Liste äquivalenter Terme wie Synonymringe oder bevorzugter Terme wie Synonymlisten, Sch Normdateien) und

- strukturierten kontrollierte Vokabularen (z.B. hierarchisch strukturierte Vokabulare wie Taxonomien, Klassifikationssysteme, Systematiken oder Thesauri)

Dabei lässt sich methodologisch eine zunehmende Aussagekraft über Artefakte und Relationen mit der Modellierung von einfachen Wortlisten hin zu komplexer gewinnen[15].

Einen großen Vorteil bietet die Linked Open Data Initiative[16], welche den Austausch und die Verknüpfung von solchen kontrollierten Vokabularen zum Ziel ha Austausch ist hierbei nicht zu vernachlässigen: Kontrollierte Vokabularien gewinnen erst dadurch ihren Nutzen, dass sie zwischen mehreren WissenschaftlerInn und so als Standard fungieren.

Dateiformate

Die Welt der Dateiformate ist eine vielfältige und komplexe: Gemäß der Objekt-Abstraktion von nestor, welche Objekte in physische, logische und konzeptionel handelt es sich bei Dateien, die Formatstandards gehorchen, um Objekte logischer Natur[17]. Das heißt, dass es sich nicht nur um physische Binärströme (Eine F Signalen auf einem Datenträger) mit einem Anfang und einem Ende handelt, sondern dass es auch Informationen gibt, welche dem Computer mitteilen, mit welc diese Binärströme erstellt wurden, mit welcher sie zu öffnen sind und wie sie im Dateisystem organisiert sind.

Error creating thumbnail: File missing

Binäre Daten. Quelle: wikimedia.org. Lizenz: GNU Free Documentation License

Solche reichlich informatiklastigen Überlegungen sind für digital arbeitende GeisteswissenschaftlerInnen wegen mehrerer Aspekte für die eigene Arbeit interess

- Die Interoperabilität von Dateiformaten mit verschiedener Software und damit auch ggf. unterschiedlichen Fragestellungen ist ein durchaus hinreichender Beantwortung einer Forschungsfrage zum Beispiel zu prüfen, ob auch andere Programme / Softwaredistributionen mit einem Dateiformat arbeiten könner verstehen können.

- Es ist auch wenig sinnvoll, ein kaum dokumentiertes, nirgendwo sonst verwendbares Dateiformat einzusetzen, wenn – wie in den Digital Humanities vorr reger Austausch mit der Community, ggf. auch über die eigenen Disziplingrenzen hinaus, statt finden soll.

- Daneben ergeben die Probleme der Langzeitspeicherung und des Langzeitzugriff gewisse Implikationen zur Wahl eines Dateiformats (Vgl. Kapitel Langz

Die in den Digital Humanities verwendeten Tools und Softwarelösungen sind sehr heterogen, das betrifft sowohl ihre Komplexität als auch ihre nur schwer mess und Beliebtheit in verschiedenen Communities

Wenn man bedenkt, dass es nur Schätzungen über die aktuell existierende Anzahl von Dateiformaten auf der Welt gibt und dass womöglich täglich neue hinzuk diese u.U. jeweils nur von einem Softwarehersteller zur Speicherung der jeweils nur in seiner Software gebräuchlichen Funktionalität erfunden werden, bedarf e gewissen Grundverständnis und einiger Kriterien, um sich in dieser Welt zurecht zu finden und sinnvolle Entscheidungen für oder gegen die Speicherung der eig einem bestimmten Format zu treffen. WissenschaftlerInnen benötigen sehr gute - meist inhaltliche - Gründe, warum sie ein singulär vorhandenes Dateiformat, w anderen Software unterstützt wird und sich auch innerhalb einer wissenschaftlichen Community keiner Bekanntheit erfreut, verwenden und sollten diese Wahl z reflektieren und kommunizieren.

Eine Übersicht über relevante Dateiformate und Metadatenstandards für die Geisteswissenschaften wurde sowohl im IANUS Projekt[18] als auch in DARIAH[1 beide sind öffentlich online zugänglich.

Zur Vergänglichkeit von Bits: Archivierung und Zugriffssicherung von Daten

Im Falle analoger Quellen und Forschungsdaten ist bekannt, dass diese von Verfall betroffen sind und mit welchen Verfallszeiträumen zu rechnen ist. So haben S naturgemäß und wenn sie nicht ständigen Kriegen oder Witterungen ausgesetzt sind, eine sehr lange Haltbarkeit – ggf. über mehrere tausend Jahre. Auch Microf Haltbarkeit von bis zu 500 Jahren bescheinigt. Neuere Datenträger, wie CD-ROMs sind hingegen von einem viel schnelleren Verfall betroffen – hier ist die Rede [20]

Es ist also nicht verwunderlich, dass die Haltbarkeit digitaler Daten eine fragile Angelegenheit ist. Beispiele aus der Praxis belegen dies:

> „The University of Southern California's neurobiologists couldn't read magnetic tapes from the 1976 Viking landings on Mars. With the data in an unknow he had to track down printouts and hire students to retype everything. 'All the programmers had died or left NASA', Miller said. 'It was hopeless to try to g the original tapes."'
>
> – A Digital Dark Age?[21]

Vint Cerf, der Mitentwickler des TCP/IP Protokolls, eines Standards mit großer Bedeutung für das Internet, sagte in Newsweek[22]:

> „People think by digitizing photographs, maps, we have preserved them forever, […] but we've only preserved them forever if we can continue to read the encode them."

Error creating thumbnail: File missing

Abbildung einer Festplatte. Lizenz: CC0 Public Domain

Die genannten Aussagen illustrieren, dass hier einige Fragen beantwortet werden müssen: Es reicht nicht, qualitativ hochwertige Daten zu generieren, sie müsse und abgelegt sein, dass auf sie auch nach längeren Zeiträumen zugegriffen werden kann und das sie durch aktuelle Hard- & Software interpretiert werden könne "gelesen" werden können, dass sie von Menschen "verstanden" werden.

Was ist Langzeitarchivierung (LZA)?

Der Begriff der *Langzeitarchivierung* (LZA) bezieht sich sowohl auf die Haltbarkeit der Datenträger, auf denen Daten gespeichert werden, als auch auf die Haltb Dateien selbst. Die Erhaltung der dauerhaften Verfügbarkeit von Informationen ist ein wichtiges Ziel: Erst mit einer gelungenen Langzeitarchivierung lassen sich Forschungsdaten langfristig auch von anderen Wissenschaftlern auswerten und nachnutzen.

Mit dem Ziel einer dauerhaften Verfügbarkeit sind einige typische Herausforderungen verbunden: Jeder kennt das Phänomen, dass es bei der Dateiübertragung, einem Videostream, bei unzuverlässiger Datenleitung zu Bitfehlern und damit auch Darstellungsfehlern in einer Datei kommen kann. Auch sind die Dateiformat deren Standardisierung und Normalisierung eine Kernkomponente bei der Pflege (englisch: Curation) von Daten – beispielsweise durch Bibliothekare und Arch International Journal of Digital Curation stellt folgende Tabelle als Übersicht über die Gefahren der Langzeitarchivierung bereit[23]:

Error creating thumbnail: File missing

Gefahren für Bits. Vgl. The International Journal of Digital Curation. Issue 1, Volume 5. 20 S. 9

Wir unterscheiden also zwischen verschiedenen Gefahren bei der langfristigen Ablage von Daten:

1. Hardware-Korruption – Die Beschädigung von Hardware-Speichern (Festplatten, DVDs etc.) durch äußere Einflüsse (Stichwort Kölner Stadtarchiv) oder Verfall
2. File-Korruption – Die Beschädigung von Dateien, wenn einzelne Bits nicht mehr lesbar sind durch entweder fehlerhafte Dateiübertragung oder beschädig
3. Format Obsoleszenz – Die Überalterung eines Dateiformats, wenn ein Dateiformat nicht weiter entwickelt wurde und von keiner aktuellen Software interp kann, gilt es als obsolet – eine langfristige Sicherung mit Gewährleistung der Lesbarkeit kann nicht mehr garantiert werden.
4. Hardware Obsoleszenz – Auch Hardware kann veralten. Man denke an die Floppy Disk. Eine Datensicherung auf Floppy Disks würde nach heutigen Maß eine ausreichende Maßnahme zur Langzeitarchivierung gelten.

Technische Lösungsstrategien und bestehende Infrastrukturangebote für die Archivierung von Daten

Die Forschung zur digitalen Langzeitarchivierung kennt folgende Ansätze, um diesen Gefahren zu begegnen:

1. Hardwaremigration – Die Migration auf dem Gebiet von Hardware meint das regelmäßige Kopieren von Daten zwischen Datenträgern. Es wird also in reg Abständen die Aktualität und Qualität der verwendeten Hardware (häufig Serverarchitekturen in Rechenzentren) geprüft und gegebenenfalls gegen aktuell ausgestauscht. Hernach ist immer ein Kopieren der enthaltenen Daten von einem zum anderen System notwendig.
2. Redundante Speicherung – Redundante Speicherung ist eine weitere Voraussetzung, um eine sichere Ablage gewährleisten. So ist eine einzelne Datei ohn Kopien an einem anderen Ort immer ein Risiko: Wenn ausgerechnet diese eine Kopie auf einem Server liegt, der einen Wasserschaden nicht überlebt hat, immer vernichtet oder kann nur durch aufwendige Maßnahmen wieder hergestellt werden – Sind hingegen weitere Kopien im Umlauf, kann auf diese aus werden. Daher empfiehlt es sich für einen Anbieter von Diensten digitaler Langzeitarchivierung, mehrere Hardware-Systeme parallel im Einsatz zu halten an unterschiedlichen Orten aufzustellen. Eine Software, die die darauf gespeicherten Daten regelmäßig überprüft und miteinander vergleicht, ggf. auch ko durch vollständige austauscht, kann hier helfen.
3. Formatmigration – Als Formatmigration wird der Vorgang, der sonst häufig Formatkonvertierung genannt wird, bezeichnet. Wenn also eine aufbewahrung ein Dateiformat besitzt, von welchem bekannt ist, dass es vermutlich nicht mehr aktuell ist, so sollte diese Datei in ein geeigneteres Dateiformat konvertie werden. Auf dem Gebiet der unterschiedlichen Medientypen existieren dabei unterschiedlich große Empfehlungen und Herangehensweisen: Für klassisch (pixelbasiert – nicht vektorbasiert), wird klassischerweise in das Dateiformat TIFF als langzeitarchivierungssicheres Dateiformat migriert. Für Videodaten einheitliche Medientypen, wie Datenbanken, existieren hingegen keine einheitlichen Überlegungen, geschweige denn einheitliche Empfehlungen.
4. Software-Emulation – Software-Emulation bezeichnet die Strategie, ein veraltetes Computer-Programm, welches nicht mehr auf aktuellen Betriebssystem nicht mehr unterstützt oder weiter entwickelt wird, zu "emulieren", d.h. nachzubilden - häufig, indem die Betriebssystemarchitektur dieser Zeit nachgebild Beispielsweise kann so CorelDraw aus den späten 90er Jahren auf einem aktuellen Apple System, bspw. MacOS X 10.9, wieder ausgeführt werden. Es ha aber um eine sehr aufwendige und experimentelle Strategie, die nur in Ausnahmefällen Anwendung findet. Am Bekanntesten sind wohl Nachbildungen vo aus den 80er Jahren, von denen mittlerweile viele per Emulation in Webtools gespielt werden können (Zum Beispiel Arcade Games aus den 80er Jahren: http://www.tripletsandus.com/80s/80s_games/arcade.htm).
5. Dokumentation – Diese Strategie wird als Ergänzung zu den vorherigen verwendet: Durch aktive und umfassende Extraktion von technischen Metadaten intensiven bibliographischen Beschreibung des Inhalts (Deskriptive Metadaten), kann sowohl das Auffinden von Daten als auch das Finden einer geeignet Software zur Interpretation erleichtert werden.

Die Punkte Hardwaremigration und redundante Speicherung sind mittlerweile hinreichend bekannt, auch Praxis moderner Rechenzentren. Für die letzten Punkte einiger Forschungs- bzw. Implementierungsbedarf. Daher müssten WissenschaftlerInnen selbst diese Punkte zumindest im Auge behalten, indem beispielsweise bei der Verwendung von Software und Dateiformaten eingehalten werden oder auch aktiv am Ende eines Forschungsprojekts in empfohlene Dateiformate konve

Die WissGrid-Initiative versucht die Lösung konzeptionell anzugehen und trifft auf Basis der Unterscheidung zwischen verschiedenen Arten des Objektbegriffs[Arten der Gewährleistung von "Speicherung". Demnach sind Langzeitarchivierungsstrategien dann erfolgreich, wenn sie folgende Ebenen berücksichtigen:

- "der physikalischen Ebene (digitale Objekte werden auf physikalischen Medien gespeichert),
- der logisch-technischen Ebene (digitale Objekte werden in bestimmten Formaten kodiert) und
- der intellektuellen Ebene (digitale Objekte erfüllen einen bestimmten Sinn für Menschen)."

Für WissenschaftlerInnen bedeutet dies, dass ihre Expertise und Mitarbeit gerade zur Erhaltung der intellektuellen Ebene von digitalen Forschungsdaten geforde geeignete Langzeitarchivierungsstrategie möglichst in Kooperation mit dem technischen Dienst, der diese umsetzen soll, abgesprochen werden muss.

Weitere bedenkenswerte Aspekte im Bezug auf die Verbreitung und Veröffentlichung von Daten

Zitierbarkeit

Im Gegensatz zu textbasierten Publikationen gibt es für Forschungsdaten – und zwar zumeist auch in den Naturwissenschaften – keine standardisierte Methode werden soll. Zur Nachvollziehbarkeit und ggf. auch Wiederholbarkeit eines digitalen Forschungsprojekts ist aber der Zugriff auf diese zugrunde liegenden Date

Es wird gemeinhin die Verwendung von persistenten Identifikatoren empfohlen.[25] Bei diesen handelt es sich um eindeutige, dauerhaft auf eine Ressource verw Zeichenketten (Also Folgen von Zahlen und Buchstaben).

So können Links und Verweise in einer Publikation langfristig zugreifbar bleiben und behalten auch nach – beispielsweise – Technologiebrüchen, Serverumzüge Firmenübernahmen weiterhin ihre Gültigkeit, da nur die *Uniform Resource Locator* (URL) hinter einem persistenten Identifikator ausgetauscht wird, nicht aber selbst. Auf diese Art ist eine dauerhafte Verfügbarkeit und ein dauerhafter Zugriff auf Daten auch im Kontext der Langzeitarchivierung sicher gestellt. Vertiefend zu persistenten Identifikatoren finden sich im Kapitel zu digitalen Infrastrukturen.

Trust

Ein weiterer – noch nicht hinreichend standardisierter – Aspekt ist der Kontext der Vertrauenssicherheit, im englischen gebräuchlicher: Trust.

Hier handelt es sich um Verfahren, um sicher zu stellen, dass einem Langzeitarchivierungssystem / einem Datenrepositorium auch vertraut werden kann. Diese v ein Audit, d.h. als eine Art Betriebsprüfung, durchgeführt. Beispiele hierfür sind das *Data Seal of Approval* (DSA)[26] oder die Norm "Audit and certification of digital repositories (ISO 16363)" von der *International Organization for Standardization* (ISO)[27]. Keiner der genannten Lösungsvorschläge wurde speziell auf von heterogenen geisteswissenschaftlichen Daten angepasst, daher bedarf es hier einer genaueren Prüfung, welche der Ansätze die passendste Lösung darstellt.

Archivierung vs. Nachnutzbarkeit

Es sollte darauf hingewiesen werden, dass die genannten Lösungsstrategien aus dem Bibliotheks- und Archivbereich stammen. Dabei wird der Fokus auf die Sic Zugriffs und der Lesbarkeit von Dateien gelegt. Ein weiteres – noch nicht ausreichend erforschtes – Feld ist deswegen die Sicherstellung ihrer Nachnutzbarkeit. sich nicht zwangsläufig um die gleiche Problemstellung: So wird im Bereich Textmedien gerne das Dateiformat PDF/A zur Langzeitarchivierung empfohlen – d nicht hinreichend berücksichtigt, dass das Dateiformat "*Portable Document Format*" (PDF) keinerlei Editieren und damit Arbeiten mit den Textdaten ermöglich lediglich um ein Dateiformat zur plattformunabhängigen Darstellung von (mehrseitigen) Text- und Bilddateien. Gerade zur Gewährleistung der Nachnutzbarkeit im Rahmen eines Forschungsdatenzyklus ein zentrales Anliegen ist, kann das Dateiformat PDF also zu einer massiven Nutzungseinschränkung führen. Hier emp alternative Dateiformate: Zum einen kann der Open Document (ODF) Standard für Office-Dokumente verwendet werden, für klassisch geisteswissenschaftliche kann der TEI Standard und die darin angebotenen Anpassungen (en: Customizations) eine gute Alternative sein. Beide Empfehlungen gelten jedoch unter Vorbe weitere Verwendungszwecke von textbasierten Informationen möglich sind, welche von den genannten Empfehlungen nur unzureichend unterstützt werden.

Insgesamt wird interessierten WissenschaftlerInnen für alle genannten Aspekte der Langzeitproblematik empfohlen, immer Nutzen und Aufwand zwischen den Optionen bei der Wahl von Dateiformaten abzuwägen. Das folgende Kapitel nennt dabei die wichtigsten Punkte, die es zu bedenken gilt.

Handlungsbedarf und offene Forschungsfragen in der Langzeitarchivierung

Das Problemfeld der Langzeitarchivierung stellt sowohl eine Herausforderung an Bibliothekare und Archivare aber auch an jeden Forschenden dar. Aus den obe Aspekten ergibt sich schnell, dass hier kein einheitlicher Lösungsvorschlag unterbreitet werden kann. Es können aber einige Bereiche identifiziert werden, in de WissenschaftlerInnen selbst Lösungsmöglichkeiten beeinflussen können:

- Eine unvollständige Liste von Dateiformaten zählt bei Wikipedia aktuell 1316 gebräuchliche Dateiformate[28]. Diese unterscheiden sich augenscheinlich i Aspekten – sowohl was ihren Anwendungsbezug als auch ihre Dokumentationstiefe, ihre Verbreitung und weiteres betrifft. Hier kann die Library of Cong Empfehlungen geben.[29] Durch die Wahl des Dateiformats lässt sich das Problem der Formatobsoleszenz zumindest eingrenzen. Kriterien, wie die Verbre Dateiformats, seine Lizenzierung und Akzeptanz über eine Community hinaus sind ein wichtiger Maßstab, um dessen Langzeitarchivierungstauglichkeit b können.
- Sollte es sich bei den desiderierten Forschungsdaten um Solche handeln, welche die Digitalisierung noch vor sich haben, sind unbedingt die Empfehlunge des Wissenschaftsrats zu beachten. Generell ist es eine gute Idee, nicht unbedingt eigene Konventionen festzulegen, sondern bereits bestehende zu übernel
- Die Frage nach der Datenablage nach Projektabschluss, sollte möglichst schon im Forschungsantrag beantwortet werden. Die verschiedenen Forschungsförderungsorganisationen machen gegebenenfalls auch Angaben zu erwünschtem Umfang und Form der abschließenden Aufbewahrung. Gene empfehlen, Forschungsdaten nicht nur lokal zu speichern, sondern Angebote des eigenen Rechenzentrums, der lokalen Hochschulbibliothek, aber auch lar bundesweiter wissenschaftlicher Infrastrukturen anzunehmen. Hier gilt es auf die oben aufgeführten Kriterien zu achten: Werden die Daten redundant ges Strategien zur Formaterkennung und -dokumentation angeboten?
- Eine hinreichend umfangreiche deskriptive Beschreibung der eigenen Forschungsdaten in einem dafür vorgesehenen und verbreiteten Metadatenstandard empfehlen. Solche Metadaten können wertvolle Zusatzinformationen liefern und erleichtern das Auffinden und die Verknüpfung zu anderen Daten in eine bieten sich die Standards der Library of Congress an.[31]

Eine nicht vollständige aber umfangreiche Liste von möglichen Kriterien, die es hinsichtlich Langzeitarchivierungsfähigkeit und Nachnutzbarkeit der eigenen F beachten gilt wurde außerdem im Rahmen von DARIAH-DE entwickelt. [32]

Links und Literatur

- Andorfer, Peter: "Forschen und Forschungsdaten in den Geisteswissenschaften. Zwischenbericht einer Interviewreihe". DARIAH-DE Working Papers Nr. DARIAH-DE, 2015 URN:urn:nbn:de:gbv:7-dariah-2015-3-8
- Bruce und Hillmann: The Continuum of METADATA Quality: Defining,Expressing, Exploiting, Published in “Metadata in Practice,” ALA Editions, 2004 Gasser, L. (2008). Value based metadata quality assessment. Library & Information Science Research, 30(1), 67-74. http://dx.doi.org/10.1016/j.lisr.2007.0
- Dangerfield, Marie-Claire; Kalshoven, Lisette (Edn.): Report and Recommendations from the Task Force on Metadata Quality, 2015, http://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf
- Kindling, Maxi (2013) Qualitätssicherung im Umgang mit digitalen Forschungsdaten. In: Information: Wissenschaft und Praxis, 64(2/3):137-147
- Neuroth, Heike, Karsten Huth, Achim Oßwald, Regine Scheffel, and Stefan Strathmann (Hg.). Eine kleine Enzyklopädie der digitalen Langzeitarchivierur nestor, 2010. http://www.nestor.sub.uni-goettingen.de/handbuch/index.php. Nestor 2010, Kap 9.1, S. 4
- Simukovic, Elena; Thiele, Raphael; Struck, Alexander; Kindling, Maxi; Schirmbacher, Peter (2014): Was sind Ihre Forschungsdaten? Interviews mit Wiss Humboldt-Universität zu Berlin. Bericht, Version 1.0. Online verfügbar unter: urn:nbn:de:kobv:11-100224755
- Tonkin, Emma. “Persistent Identifiers: Considering the Options.” Ariadne, no. 56 (2008). http://www.ariadne.ac.uk/issue56/tonkin
- http://linkeddata.org/
- http://www.europeana.eu
- http://www.re3data.org/
- http://www.ianus-fdz.de/it-empfehlungen/dateiformate
- http://datasealofapproval.org/en/

Anmerkungen

1. ↑ Für Digitalisate und darauf basierenden Forschungsergebnissen, ist es essentiell festzustellen, inwiefern die Rekonstruktion dem historischen Objekt nah welche Erkenntnisse belegbar sind.
2. ↑ In einem Essay von Ricky Erway von OCLC werden digital erzeugte Daten als Daten definiert, die digital erstellt wurden und in digitaler Form bearbeit (Erway, Ricky: Defining “Born Digital”. An Essay by Ricky Erway, OCLC Research, 2010)
3. ↑ Weitere Datenformate und Beispiele wurden in durchgeführten Interviews mit Fachwissenschaftlern an verschiedenen deutschen Universitäten und Forschungseinrichtungen, z.B. Simukovic, Elena; Thiele, Raphael; Struck, Alexander; Kindling, Maxi; Schirmbacher, Peter (2014): Was sind Ihre Forschu Interviews mit Wissenschaftlern der Humboldt-Universität zu Berlin. Bericht, Version 1.0. Online verfügbar unter: urn:nbn:de:kobv:11-100224755 oder P "Forschen und Forschungsdaten in den Geisteswissenschaften. Zwischenbericht einer Interviewreihe". DARIAH-DE Working Papers Nr. 10. Göttingen: D 2015 URN:urn:nbn:de:gbv:7-dariah-2015-3-8)
4. ↑ Johanna Puhl, Peter Andorfer, Mareike Höckendorff, Stefan Schmunk, Juliane Stiller, Klaus Thoden: "Diskussion und Definition eines Research Data L digitalen Geisteswissenschaften".*DARIAH-DE Working Papers*Nr. 11. Göttingen: DARIAH-DE, 2015 URN:**urn:nbn:de:gbv:7-dariah-2015-4-4** http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2015-11.pdf
5. ↑ Peter Andorfer: "Forschen und Forschungsdaten in den Geisteswissenschaften. Zwischenbericht einer Interviewreihe". *DARIAH-DE Working Papers*Nr. DARIAH-DE, 2015 URN:urn:nbn:de:gbv:7-dariah-2015-3-8
6. ↑ Johanna Puhl, Peter Andorfer, Mareike Höckendorff, Stefan Schmunk, Juliane Stiller, Klaus Thoden: "Diskussion und Definition eines Research Data L digitalen Geisteswissenschaften".*DARIAH-DE Working Papers*Nr. 11. Göttingen: DARIAH-DE, 2015 URN:**urn:nbn:de:gbv:7-dariah-2015-4-4** http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2015-11.pdf
7. ↑ Aus: Fachspezifische Empfehlungen für Daten und Metadaten, http://dev2.dariah.eu/wiki/pages/viewpage.action?pageId=20058160
8. ↑ www.europeana.eu
9. ↑ Dangerfield, Marie-Claire; Kalshoven, Lisette (Edn.): Report and Recommendations from the Task Force on Metadata Quality, 2015, http://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf
10. ↑ Weiterführende Literatur zur Metadatenqualität in digitalen Bibliotheken: Bruce und Hillmann: The Continuum of METADATA Quality: Defining,Expr Exploiting, Published in “*Metadata in Practice*,” ALA Editions, 2004,Stvilia, B., & Gasser, L. (2008). Value based metadata quality assessment. Library & Science Research, 30(1), 67-74. http://dx.doi.org/10.1016/j.lisr.2007.06.006 & Park, Jung-Ran. “Metadata Quality in Digital Repositories: A Survey of the the Art.”*Cataloging & Classification Quarterly*47, no. 3–4 (April 9, 2009): 213–28. doi:10.1080/01639370902737240.
11. ↑ Alex, B. and Burns, J. 2014. Estimating and Rating the Quality of Optically Character Recognised Text. In Proceedings of DATeCH 2014, Madrid, Spai
12. ↑ siehe die ergänzte und aktualisierte Denkschrift "Sicherung guter wissenschaftlicher Praxis" der DFG, 2013, http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf

13. ↑ Kindling, Maxi (2013) Qualitätssicherung im Umgang mit digitalen Forschungsdaten. In: Information: Wissenschaft und Praxis, 64(2/3):137-147
14. ↑ http://www.re3data.org/
15. ↑ Für weitere Informationen: https://dev2.dariah.eu/wiki/display/publicde/5.+Kontrolliert-Strukturierte+Vokabulare
16. ↑ Vgl. http://linkeddata.org/
17. ↑ Neuroth, Heike, Karsten Huth, Achim Oßwald, Regine Scheffel, and Stefan Strathmann (Hg.). Eine kleine Enzyklopädie der digitalen Langzeitarchivier nestor, 2010. http://www.nestor.sub.uni-goettingen.de/handbuch/index.php. Nestor 2010, Kap 9.1, S. 4
18. ↑ http://www.ianus-fdz.de/it-empfehlungen/dateiformate
19. ↑ https://dev2.dariah.eu/wiki/pages/viewpage.action?pageId=38080370
20. ↑ http://wp.ub.hsu-hh.de/13800/haetten-sies-gewusst-geschichte-speichermedien-begann-40-000-v-chr/
21. ↑ Coming Soon: A Digital Dark Age?. 2013. http://www.cbsnews.com/news/coming-soon-a-digital-dark-age/
22. ↑ http://www.newsweek.com/2015/07/03/storing-digital-data-eternity-345557.html
23. ↑ http://ijdc.net/index.php/ijdc/article/view/143/205, S.9
24. ↑ Neuroth, Heike, Karsten Huth, Achim Oßwald, Regine Scheffel, and Stefan Strathmann (Hg.). Eine kleine Enzyklopädie der digitalen Langzeitarchivier nestor, 2010. http://www.nestor.sub.uni-goettingen.de/handbuch/index.php
25. ↑ Tonkin, Emma. "Persistent Identifiers: Considering the Options." Ariadne, no. 56 (2008). http://www.ariadne.ac.uk/issue56/tonkin
26. ↑ http://datasealofapproval.org/en/
27. ↑ Vgl. UK DATA ARCHIVE: HOW TO CURATE DATASTANDARDS OF TRUST. http://www.data-archive.ac.uk/curate/trusted-digital-repositories/star
28. ↑ https://en.wikipedia.org/wiki/List_of_file_formats
29. ↑ http://www.loc.gov/preservation/resources/rfs/TOC.html
30. ↑ http://www.dfg.de/formulare/12_151/12_151_de.pdf
31. ↑ http://www.loc.gov/standards/
32. ↑ https://dev2.dariah.eu/wiki/pages/viewpage.action?pageId=38080370

Alles was Recht ist: Urheberrecht und Lizenzierung von Forschungsdaten

Unter dem Schlagwort Digital Humanities finden neue Techniken der quantitativen Analyse vermehrt Eingang in den Methodenkanon der Geistes- und Kulturw dies erfordert nun auch ein Umdenken in der Verwaltung von Zugriffs- und Nutzungsrechten von Forschungsdaten. Im Zuge dieser Umwälzungen ist die bisheri auf den Einzelfall bezogene Klärung von rechtlichen Aspekten zur Nachnutzung von Daten aufgrund ihrer Masse nicht mehr praktikabel. Die Verwendung von S bietet sich hier als Lösung an. Doch stellen sich im Zusammenhang mit rechtlichen Aspekten von Forschungsdaten noch weitere Fragen: Wem gehören die Fors eigentlich? Welche Rechte kann der/die WissenschaftlerIn geltend machen? Welche Rechte hat die arbeitgebende Institution an Forschungsdaten? Sind Daten üb geschützt?

Nachnutzung fremder Inhalte in der wissenschaftlichen Arbeit

Anschauen oder Lesen von Werken ist rechtlich immer erlaubt, dabei handelt es sich nicht um ein Nutzungsrecht im Sinne des Urheberrechtsgesetzes, sondern u sogenannten Werkgenuss, der rechtlich nicht reglementiert ist. Ob es sich beim *Text and Data Mining* (TDM) aber um einen technisierten Werkgenuss handelt, i Spätestens wenn die Daten des Text und Data Minings jedoch aus Gründen der Transparenz und der Überprüfbarkeit der Forschungsergebnisse veröffentlicht we dies nicht ohne eine Vervielfältigung der analysierten Daten und ist damit nicht mehr lizenzfrei. Für die Wissenschaft und Forschung gibt es jedoch einige rechtl Sonderbestimmungen, bedeutsam sind hier das Zitatrecht und wissenschaftliche Schranken, die ein Remixing und kollaborative Bearbeitungen erlauben oder au Inhalten für Unterricht und Forschung im begrenzten Umfang ermöglichen[1].

Rechte der/des Datenproduzenten und der arbeitgebenden Institution

WissenschaftlerInnen im Anstellungsverhältnis, die Forschungsdaten oder Publikationen - also Werke - erstellen, haben als Urheber das Recht auf Namensnennu Urheberpersönlichkeitsrecht ist nicht übertragbar, somit bleibt das Recht auf Namensnennung den WissenschaftlerInnen immer erhalten. Die Leistungsschutzrechte/Nutzungsrechte für Vervielfältigung und Verbreitung fallen in den meisten Fällen an die arbeitgebende Institution, da das Erstellen von W üblichen Aufgaben im Rahmen eines Dienst- oder Anstellungsverhältnis zählt und die Ergebnisse somit der Institution gehören.[2]

Offene Daten und Standardlizenzen

Open Data steht für einen kulturellen Wandel im Verhältnis von BürgerInnen und Staat, der zu mehr Transparenz, mehr Teilhabe und einer intensiveren Zusamm Öffentlichkeit führen kann. Dieses Konzept ist im akademischen Bereich nicht neu und ähnelt den Konzepten von Open Access, Open Content und Open Source Institutionen, die durch Steuergelder finanziert werden, produzieren enorme Datenmengen, zu denen u.a. statistische Daten, Forschungsdaten und Kulturdaten g Daten als offene Daten vor, können sie von BürgerInnen, Nichtregierungsorganisationen, Bildungseinrichtungen, JournalistInnen und Firmen auf vielfältige Wei werden.

Wie Offenheit definiert ist und wie man in dem Zusammenhang die größtmögliche Zugänglichkeit und Nutzbarkeit digitaler Informationen sicherstellen kann, is Definition der Open Knowledge Foundation festgehalten: „Digitale Daten und Inhalte sind dann offen, wenn sie von allen gleichermaßen frei genutzt, kombinie weiterverbreitet werden können – maximal eingeschränkt durch die Pflicht der Namensnennung und/oder der Weitergabe unter gleichen Bedingungen"[3].

Zu den wichtigsten Kriterien offener Daten zählen:

Verfügbarkeit und Zugang: Das digitale Werk soll als Ganzes verfügbar sein, zu Kosten, die nicht höher als die Reproduktionskosten sind, vorzugsweise zum Download im Internet. Das Werk soll ebenso in einer zweckmäßigen und modifizierbaren Form verfügbar sein – und zwar ohne notwendigen Login, rund um di abrufbar.

Wiederverwendung und Nachnutzung: Die Daten müssen unter denjenigen Bedingungen bereitgestellt werden, die die Wiederverwendung, Nachnutzung und anderen Datensätzen erlauben. Die Daten müssen maschinenlesbar sein, damit sie von z.B. Entwickler/innen oder Datenjournalist/innen verarbeitet werden könr

Universelle Beteiligung: Jede Person muss in der Lage sein, die Daten zu nutzen, wiederzuverwenden und weiterzugeben. Es darf keine Diskriminierung gegen Personen oder Gruppen vorliegen. Die Nachnutzung darf also nicht auf einzelne Bereiche begrenzt werden (z.B. nur in der Bildung), noch dürfen bestimmte Nu für kommerzielle Zwecke) ausgeschlossen sein[4].

Sind Daten erst einmal offen, können sie von verschiedenen AkteurInnen wie Software-EntwicklernInnen, DatenjournalistInnen, WissenschaftlerInnen etc. auch Institutionsgrenzen hinaus genutzt werden. Und wenn diese verarbeitet werden, nützen sie auch der Gesellschaft.

Neue Möglichkeiten durch alternative Lizenzierungen

> „Lizenzen machen halt vor Grenzen, offene Daten jedoch nicht."
> – Jörg Prante[5]

Sprechen wir von der digitalen Zeit - so sprechen wir automatisch über Daten. Damit Personen außerhalb der eigenen Institution Inhalte wie Bilder, Videos, Ton und die dazugehören beschreibenden (Meta)-daten frei nachnutzen können, müssen diese unter einer offenen Lizenz stehen. Dafür braucht die Institution, die di machen will, ein zeitlich und räumlich unbeschränktes Nutzungsrecht des Werkes, das alle bekannten und zum gegebenen Zeitpunkt noch unbekannte Nutzungs Sonst gilt automatisch das Urheberrecht mit seinen Nutzungseinschränkungen. Urheberrechtlicher Schutz entsteht automatisch mit der Schaffung eines Werkes.

Offene Lizenzen erlauben das Teilen von Informationen und geben DatennutzerInnen mehr Freiraum für die Umsetzung von Projekten sowie Ideen. Sie ermögli abgestufte Wahrnehmung von Rechten. Vom urheberrechtlichen Standard des „Alle Rechte vorbehalten“ hin zu „Manche Rechte vorbehalten“ der *Creative Com* Lizenzen bzw. „Keine Rechte vorbehalten“ für den Bereich der Public Domain.

Zu den bekanntesten offenen Lizenzen gehören *GNU General Public License* (GNU GPL) für Open Source Software sowie Creative Commons (kreative Allme urheberrechtlich schützbaren Werke.

Wer oder was ist Creative Commons?

Creative Commons (CC) ist sowohl eine amerikanische Non-Profit-Organisation, als auch eine internationales Netzwerk von JuristInnen, AktivistInnen und Kre Commons-Lizenzen basieren auf dem bestehenden Urheberrecht und stellen den Versuch dar, sich dem Ideal einer Wissens- oder Kreativallmende in Form eines großen Pools an alternativ lizenzierten Werken anzunähern. RechteinhaberInnen, die Werke unter eine Creative Commons-Lizenz stellen, räumen Dritten bestim Nutzungsmöglichkeiten ein, die sonst – ohne eine aktive Lizenzierung - vorbehalten blieben. Gleichzeitig berücksichtigen die Creative Commons-Lizenzen stärk unterschiedlichen Interessen der UrheberInnen, die entscheiden, wie andere ihre Werke nutzen, weitergeben und verwerten können.

Die folgende Tabelle zeigt, welche offenen Lizenzen gemäß der Open Definition für welche Datenarten verwendet werden.[6]

Lizenz	Erklärung
CC-BY Namensnennung	Neben dem Hinweis auf den Autor, die Quelle, Rechteinhaber und die Lizenz enthält diese CC-Variante keine weiteren Einschränkungen fü ist in der Verwendung des Werkes frei und kann es in jeder erdenklichen Form bearbeiten, verbreiten, verbessern und darauf aufzubauen, au verwerten. Damit ist die Nutzung eines Werkes z.B. in Remixes oder Mashups möglich.
CC-BY-SA Namensnennung	Weitergabe unter gleichen Bedingungen: Auch diese Lizenz erlaubt sowohl die Bearbeitung eines Werkes als auch die kommerzielle Nutzu Bearbeitungen dürfen aber nur unter den gleichen oder vergleichbaren Lizenzbestimmungen veröffentlicht werden. Alle neuen Werke, die a ursprünglichen Werk aufbauen, werden unter derselben Lizenz stehen, also sind auch kommerziell nutzbar. Diese Lizenz wird oft mit “Cop im Bereich freier und Open Source Software verglichen. Der Autor, die Quelle, Rechteinhaber und die Lizenz ist anzugeben.
CC-BY-ND Namensnennung	Namensnennung, keine Bearbeitung: Der Autor ist wie in den oben genannten Lizenzen zu benennen. Diese Lizenz gestattet keine Bearbeit kommerzielle Nutzung ist hingegen erlaubt.

Error creating thumbnail: File missing

Creative Commons – Choose a License

Die genannten Lizenzarten gibt es ergänzt um das Non Commercial-Modul (nicht-kommerziell). Aus den drei oben beschriebenen Lizenztypen werden dadurch noch zusätzlich: CC-BY-NC, CC-BY-NC-SA und CC-BY-NC-ND. Die ersten drei Lizenzen räumen die größtmögliche Nutzung für Dritte ein, die Lizenzen mit Commercial und/oder Non Derivative-Vermerk schränken die Nutzungsmöglichkeiten auf unterschiedliche Weisen ein und gelten daher nicht als offen im Sinne Definition und der Definition von freedomdefined.org.

Public Domain (Gemeinfreiheit)

Error creating thumbnail: File missing

Urheberrechtsschutzfristen, Wikipedia: CC-BY 3.0 Balfour Smith, Canuckguy, Badseed. - Original image by Balfour Smith at Duke University at page (direct link). Vectorized by Badseed using BlankMap-World6 as a basemap. http://www.publicdomainday.org/sites/www.publicdomainday.eu/files/World_copyright-terms.jpg

Die sogenannte Public Domain, die Gemeinfreiheit oder Allmende, beinhaltet Werke, bei denen der urheberrechtliche Schutz abgelaufen ist bzw. Inhalte, die nie geschützt waren. Die Public Domain spielt eine wichtige gesellschaftliche und wirtschaftliche Rolle, fördert die Schöpfung und Umsetzung neuer Ideen als freie

Werke der Public Domain unterliegen keinerlei urheberrechtlicher Nutzungsbeschränkungen. Dies können zum Beispiel Ideen, Konzepte, Zahlen, Namen und T

In Deutschland und vielen anderen Ländern fallen Werke erst 70 Jahre nach dem Tod des Urhebers und der Urheberin in die Gemeinfreiheit. Die sogenannten Sc werden tendenziell verlängert und verhindern, dass die Gesellschaft von Werken der Public Domain profitiert.

Die Lizenz CC0 – public domain dedicated bildet diese Gemeinfreiheit rechtlich nach und ermöglicht es, Werke direkt in die Public Domain bedingungslos freiz der Erfinder des *World Wide Web*e (WWW), Tim Berners Lee, getan hat), sodass Dritten die maximale Nutzungsfreiheit eingeräumt wird.

Vorgehen bei der Lizenzierung

Es ist wichtig, eine Lizenz zu finden, die für die Art von Material angemessen ist, das geöffnet wird. Die Anforderung, bei der Nachnutzung eines Artikels, Gedi UrheberInnen korrekt zu benennen, ist tief verankert in den Normen der wissenschaftlichen Praxis und ist das Mittel, mit dem NutzerInnen eines Werkes im Kor nachvollziehen können, welche Teile davon ein Original sind.

Bei Daten gibt es allerdings häufig sehr gute Gründe, von der Pflicht zur Namensnennung abzusehen. Eine Anzahl prominenter Datenportale für das Kulturerbe, akzeptieren nur Daten, die unter der *Creative Commons Zero-Lizenz* (CC0) zugänglich gemacht werden. Metadaten eines Werkes sind umso nützlicher, je besser Daten kombiniert werden können (Linked Open Data). Es ist daher empfehlenswert, für Metadaten die CC0-Lizenz zu verwenden, da sonst u.a. die Kette der Na sehr lang werden kann.

Wichtig ist es, die Lizenzierung frühzeitig in den einzelnen wissenschaftlichen Arbeitsschritten mitzubedenken. Folgende Punkte sollten beachtet werden[7]:

- Integrieren Sie die Lizenzierung Ihrer Forschungsdaten in die Veröffentlichungsprozesse bzw. -richtlinien Ihrer Institution.
- Im Falle der Generierung von Forschungsdaten in einem Kooperationsprojekt, sollte bereits im Projektantrag festgelegt werden, unter welcher Lizenz die veröffentlicht werden. Die *Deutsche Forschungsgemeinschaft* (DFG) empfiehlt explizit die Verwendung von CC-BY-SA für im Open Access veröffentlich für Metadaten [8].
- Es sollten für die vollständige Lizenzierung immer folgende Informationen angegeben werden: Name des Rechteinhabers, Jahr der Veröffentlichung und c
- Der Verwendung von offenen Standardlizenzen sollte der Vorzug gegeben werden.
- Versichern Sie sich, dass Sie die Rechte an allen Daten haben, die Sie veröffentlichen wollen.
- Entscheiden Sie, ob Sie die kommerzielle Nutzung Ihrer Daten erlauben wollen.
- Bei Creative Commons Lizenzen ist zu beachten, dass sie nicht-exklusiv sind, d.h. das Inhalte neben der CC Lizenz auch unter weiteren Lizenzen stehen k sollte jedoch vermieden werden, um rechtliche Konflikte zu vermeiden.
- Berücksichtigen Sie, dass für verschiedene Teile Ihrer Datensammlung unterschiedliche Lizenzen zur Anwendung kommen können. Wählen Sie deshalb j Lizenz für Metadaten, kontrollierte Vokabulare oder Digitale Objekt/Inhalte (Bilder,Volltexte, Audiobeiträge, Videos etc.) bzw. Datenbanken und Daten D
- Prüfen Sie aktuelle juristische Handreichungen für die Geisteswissenschaften und diskutieren Sie Ihre Fragen nach Möglichkeit auch mit interessierten Kc

Links und Literatur

Alex Ball. 2012. How to License Research Data. DCC How-to Guide. Edinburgh. http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_To_License_Research_Data.pdf.

Nikolaos Beer, Kristin Herold, Wibke Kolbmann, Thomas Kollatz, Matteo Romanello, Sebastian Rose, Niels-Oliver Walkowski, Felix Falko Schäfer, und Maur 2014. „Datenlizenzen für geisteswissenschaftliche Forschungsdaten: Rechtliche Bedingungen und Handlungsbedarf.“ DARIAH-DE Report, DARIAH-DE Work http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2014-4-8.

Creative Commons Deutschland, http://de.creativecommons.org/

Michael Fehling. 2014. „Verfassungskonforme Ausgestaltung von DFG-Förderbedingungen zur Open-Access-Publikation.“ *Ordnung der Wissenschaft* 4: 179–2 http://www.ordnungderwissenschaft.de/Print_2014/24_fehling_dfg_odw_ordnung_der_wissenschaft_2014.pdf.

Paul Klimpel, John H. Weitzmann: "Forschen in der digitalen Welt. Juristische Handreichung für die Geisteswissenschaften". DARIAH-DE Working Papers Nr. DARIAH-DE, 2015. URN: urn:nbn:de:gbv:7-dariah-2015-5-0 (Veröffentlichung August 2015)

J. Klump 2012. „Offener Zugang zu Forschungsdaten.“ Herausgegeben von U. Herb. Open Initiatives: Offenheit in der digitalen Welt und Wissenschaft, 45–53. http://eprints.rclis.org/handle/10760/17213.

Thinh Nguyen. 2012. „Freedom to Research: Keeping Scientific Data Open, Accessible, and Interoperable.“ http://sciencecommons.org/wp-content/uploads/free research.pdf

Anmerkungen

1. ↑ Paul Klimpel, John H. Weitzmann: "Forschen in der digitalen Welt. Juristische Handreichung für die Geisteswissenschaften". DARIAH-DE Working Pa Göttingen: DARIAH-DE, 2015, S. 10-15. URN: urn:nbn:de:gbv:7-dariah-2015-5-0 (Veröffentlichung August 2015)

2. ↑ Paul Klimpel, John H. Weitzmann: "Forschen in der digitalen Welt. Juristische Handreichung für die Geisteswissenschaften". DARIAH-DE Working Pa Göttingen: DARIAH-DE, 2015, S. 20 ff.. URN: urn:nbn:de:gbv:7-dariah-2015-5-0 (Veröffentlichung August 2015)
3. ↑ Open Knowledge Foundation (Stand: 2014): Open Definition, URL: http://opendefinition.org/od/deutsch/ (Abfrage: 22.02.2014)
4. ↑ Open Knowledge Foundation (Stand: 2015): Open Data, URL: http://okfn.de/opendata/#sthash.ESNSzIgL.dpuf (Abfrage: 22.02.2015)
5. ↑ http://open-data.fokus.fraunhofer.de/stand-der-lizenznutzung-auf-ausgewahlten-datenportalen/
6. ↑ Für weitere Informationen siehe auch: http://creativecommons.org/licenses/
7. ↑ Beer, Nikolaos, Kristin Herold, Wibke Kolbmann, Thomas Kollatz, Matteo Romanello, Sebastian Rose, Niels-Oliver Walkowski, Felix Falko Schäfer, u Heinrich. 2014. „Datenlizenzen für geisteswissenschaftliche Forschungsdaten: Rechtliche Bedingungen und Handlungsbedarf.“ DARIAH-DE Report, DA Working Papers, 6. http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2014-4-8
8. ↑ DFG (Stand: 2015):Digitalisierungsrichtlinien, http://www.dfg.de/formulare/12_151/12_151_de.pdf, S.40f. (Abfrage: 10. August 2015).

Methoden und Werkzeuge in den Digital Humanities

Vielfalt digitaler Methoden und Werkzeuge

Ein großer Vorteil digital gespeicherter Daten liegt darin, dass diese nun am Computer durchsucht, visualisiert und analysiert werden können. Hierfür steht eine entwickelter Untersuchungsmethoden und entsprechender Software-Werkzeuge zur Verfügung, die ein breites Spektrum an geisteswissenschaftlichen Diszipline teilweise sehr unterschiedliche Anforderungen an das technische Vorwissen der AnwenderInnen stellen.

Eine Übersicht über digitale Werkzeuge, die sich für bestimmte geisteswissenschaftliche Fragestellungen eignen, wurde in *Digital Research Infrastructure for th Humanities* (DARIAH) [1] sowie im *Digital Research Tools* (DIRT) -Directory [2] zusammengetragen. Beide Übersichten sind online zugänglich. Im folgenden Möglichkeiten digitaler Visualisierungs- und Analysewerkzeuge beispielhaft anhand der Raum-Zeit-Visualisierung im GeoBrowser und der Stilometrischen Tex Stylo-Paket vorgestellt werden.

Raum-Zeit Visualisierung

Gerade große Datenmengen lassen sich gut durch digitale Werkzeuge erschliessen und durch Visualisierungen analysieren. Auf diese Weise geraten auch nicht u evidente, strukturelle und inhaltliche Zusammenhänge in den Blick. Ein Beispiel für Raum-Zeit Visualisierung in den Digital Humanities ist der DARIAH-DE C

> „Unter der Visualisierung von Daten versteht man in den Digital Humanities einen computergestützten Prozess, mit dessen Hilfe geistes- und kulturwisser Daten so dargestellt und analysiert werden können, dass eine visuelle Repräsentation der inhärenten kontextualen bzw. inhaltlichen Zusammenhänge entst diese Weise können insbesondere größere Daten- und Quellenmengen analysiert werden, die von einzelnen ForscherInnen mit klassischen, nicht-digitalen nicht oder allenfalls nur mit erheblichem Zeit- und Ressourcenaufwand durchgeführt werden könnten.“
> – Kollatz, Thomas; Schmunk, Stefan: Datenvisualisierung: Geo-Browser und DigiVoy

DARIAH-DE Geo-Browser

Error creating thumbnail: File missing

Abbildung 5.1: DARIAH-DE Geo-Browser – Visualisierung von Grabmalen mit Symbolen http://steinheim-institut.de/cgi-bin/epidat

"Der Geo-Browser[4] vereint drei korrelierende Elemente: eine Karte, eine Zeitleiste sowie die Dokumentation der visualisierten Datengrundlage. Bei den Karte zwischen frei wählbaren zeitgenössischen und mehreren historisierenden Karten gewählt werden. Zudem besteht die Möglichkeit, eigenes georeferenziertes Kar einzubinden.

Das im Geo-Browser hinterlegte Standard-Kartenmaterial deckt einen Zeitraum von über 2.000 Jahren ab, sodass analog zur Periode der Datengrundlage meist e entsprechende historisierende Karte zugeschaltet werden kann. Datensets aus dem beginnenden 20. Jahrhundert etwa können auf dem entsprechenden historisier Kartenmaterial mit der flächen- und grenzgetreuen Staatenwelt am Vorabend des Ersten Weltkrieges dargestellt werden. Um die Vielzahl der Einzelorte bzw. der Datenmengen zu strukturieren, werden bei der Visualisierung im Geo-Browser Einzeldaten nach Dichte und Quantität zu regionalen Häufungen („heaps“) akkur

Die zeitliche und quantitative Dimension des Gesamtdatenbestandes je aktueller Auswahl wird in einem Graph auf der Zeitleiste dargestellt. Wird ein Punkt auf angesteuert, werden die entsprechenden Punkte der Zeitleiste sowie die Dokumentationsfelder hervorgehoben. Wird ein Zeitpunkt oder eine Zeitspanne auf der Z ausgewählt, diese bewegt oder animiert, werden stets die korrelierenden Georeferenzen und Dokumentationsfelder hervorgehoben […] Jederzeit kann aus dem a markierten Zwischenergebnis ein neues Datensample generiert werden – etwa, um Entwicklungen in unterschiedlichen Zeiträumen oder Regionen miteinander z Zur lokalen Weiterverarbeitung oder als Grundlage weiterer Visualisierungsschritte kann das Datenset auch […] exportiert werden" (Kollatz/Schmunk S. 173)

DARIAH-DE Datasheet Editor

Error creating thumbnail: File missing

Abbildung 5.2: DARIAH-DE Datasheet-Editor
http://geobrowser.de.dariah.eu/beta6/edit/

"Der Datasheet-Editor[5] bietet NutzerInnen zwei Optionen, eigene Daten für die Visualisierung im Geo-Browser aufzubereiten: zum einen den Import und die a Anreicherung bestehender CSV-Tabellen, zum anderen die Direkteingabe raum- und zeitbezogener Daten.

In der Regel werden die Datensätze direkt in den von TextGrid und DARIAH-DE gemeinsam genutzen DARIAH-DE-Storage überführt, gesichert und auch dor unabhängig davon, ob sie in den Datasheet-Editor importiert oder von Anfang an darin erstellt wurden. […] Die Option zur Direkteingabe der Daten ist sehr ein erforderlich sind lediglich Orts- und Zeitangaben. Anschließend werden Geolokalisierungen (Längen- und Breitenangaben) unter Verwendung des Getty *Thesau Geographic Names* (TGN), *Open Geo Names* (OGN) und/oder *Open Street Maps* (OSM) (semi)automatisch ergänzt – ein Verfahren, das nicht nur Zeit spart, so Ortsdaten zugleich mit den fehlenden Längen- und Breitenangaben und den entsprechenden eindeutigen Identifikatoren der Vokabulare angereichert werden.

Anschließend können die im Datasheet-Editor angereicherten Daten nicht nur im Geo-Browser visualisiert und analysiert, sondern auch in weiteren Anwendung werden. Die Visualisierung direkt aus dem Datasheet-Editor heraus ermöglicht die Direktkontrolle auf den Karten des Geo-Browsers. Sollten Orte bei der autom Georeferenzierung falsch zugeordnet werden – wie etwa bei identischen Ortsnamen in unterschiedlichen Ländern oder Regionen (z.B. Paris/Texas und Paris/Fra Frankfurt am Main und an der Oder) –, kann dies in der Nachbearbeitung leicht korrigiert werden, nämlich wiederum unter Zuhilfenahme der eingebunden Thes Alternativorte werden in einem Drop-down-Menü angezeigt, aus denen dann die korrekte Ortsangabe samt Koordinaten und Identifikator übernommen werden k

Zur Optimierung des Zugriffs und der Skalierbarkeit bei der Verarbeitung von größeren Datenmengen wird bislang auf einen TGN-Dump zugegriffen, der von T DARIAH-DE gehostet wird. Im ersten Quartal 2015 wird dies umgestellt und eine seit Sommer 2014 zugängliche Schnittstelle des TGN direkt abgefragt. Dies e Direktkontrolle auf den Karten des Geo-Browsers. Sollten Orte bei der automatischen Georeferenzierung falsch zugeordnet werden – wie etwa bei identischen O unterschiedlichen Ländern oder Regionen […], kann dies in der Nachbearbeitung leicht korrigiert werden, nämlich wiederum unter Zuhilfenahme der eingebunc Alternativorte werden in einem Drop-down-Menü angezeigt, aus denen dann die korrekte Ortsangabe samt Koordinaten und Identifikator übernommen werden k (Kollatz/Schmunk S. 171f.)

Stilometrische Textanalyse

Ein anderes großes Arbeitsfeld, dass sich mit der fortschreitenden Digitalisierung eröffnet, ist die computergestützte, quantitative Analyse digitalisierter literaris befinden sich eine ganze Reihe orginär digitaler Forschungsmethoden in der Entwicklung, die nun nicht mehr der Beschleunigung oder Erleichterung von Vorge dienen, die schon lange vorher auch ohne die Hilfe eines Computers genutzt praktiziert wurden. Neben so hilfreichen Funktionen wie einer Volltextsuche, die du Archivierungsform möglich werden, können literarische Texte nun auch mit empirisch-statistischen Verfahren untersucht werden.

Diese erlauben prinzipiell die Berücksichtigung einer weitaus größere Menge von textbasierten Daten, als man sie sonst durch Lesen oder Recherchieren erfasse allem eröffnen sie einen schnellen Blick auf die Dimensionen des Forschungsgegenstandes, die bisher kaum erfassbar waren, womit das klassische Methodenrep Philologie durch gänzlich neue Verfahren ergänzt werden kann.

Eine der häufige Anwendung der computergestützten Textanalyse in der Forschungspraxis ist die Zuschreibung eines Textes zu einem bestimmten Autor mit Hil Die Stilometrie ist ein Set statistischer Verfahren, die es erlauben, stilistische Unterschiede sichtbar und auch messbar zu machen. Sie ermöglichen es, den Stil v Autoren zu vergleichen, anonyme oder undatierte Texte einem Autor oder einer Epoche zuzuordnen oder spezifische Eigenschaften innerhalb einer Gattung hera Länger etablierte Methoden in diesem Bereich sind die *Principal Component Analysis* (PCA) und die Messung stilistischer Distanzen durch Textabstandsmaße.

Wie funktioniert Stilometrie?

Die stilometrische Forschung begann mit der Beobachtung, dass AutorInnen bestimmte Gewohnheiten und Vorlieben bei der Wahl ihres Vokabulars haben. Dies Vorlieben zeigen sich schon in den häufigsten Funktionswörtern, wie "und", "der" und "die". Ordnet man alle Wörter, die in einem Text, oder in einem ganzen T vorkommen nach ihrer Häufigkeit, so reicht oft schon die Berücksichtigung der Häufigkeiten der ersten 100 Wörter in dieser Liste, um stilistische Unterschiede herauszuarbeiten zu erkennen.

Je nach Fragestellung kann aber auch die Berücksichtigung anderer Eigenschaften der zu untersuchenden Texte sinnvoll sein, z.B. Satzlängen, die Häufigkeiten von bestimmten grammatischen Konstruktionen oder von seltenen Inhaltswörtern. Diese Eigenschaften eines Textes, die einer Analyse zu Grunde liegen, bezeic Features. Grundsätzlich kann fast jede Eigenschaft eines Textes als Feature in der Stilometrie zum Einsatz kommen – vorausgesetzt sie ist messbar und erlaubt e einen eindeutigen Wert zuzuordnen. In der Praxis, gerade bei der Autorenschaftsattribution, haben sich als die gängigsten Features tatsächlich die Häufigkeiten d Wörter etabliert.

Aber wie erkennt man nun relevante Unterschiede in einer Vielzahl von Features, z.B. in zwei Reihen von jeweils 100 Worthäufigkeiten? Der, auch in der Stilon Weg, in einer Menge von Informationen relevante Muster zu finden ist die Reduktion auf ein vereinfachendes Modell. Für die Stilanalyse werden einzelne Texte in einem mehrdimensionalen Raum modelliert. Die Dimensionen bzw. die Achsen des Koordinatensystems sind in diesem Modell die Features, die Position eine eines Textes auf einer bestimmten Achse entspricht dem Wert, den der Text für dieses Feature hat, also z.B. der Häufigkeit, mit der das entsprechende Wort in di vorkommt. Das heißt aber nun, dass ein Textkorpus, wenn nur die 100 häufigsten Wörter als Features berücksichtigt werden, als Wolke von Punkten in einem K mit 100 Dimensionen modelliert wird! Gleichzeitig lassen sich aber nur höchstens 3 Dimensionen sinnvoll graphisch abbilden. Wie soll also dieses "vereinfache helfen, relevante Muster zu erkennen? Ein etabliertes mathematisches Verfahren, mit dieser Art von Datenmodell umzugehen ist die Principal Component Analy eines der ersten Verfahren, die in der quantitativen Textanalyse eingesetzt wurden.

Strukturen erkennen im hochdimensionalen Raum: Die *Principal Component Analysis*

Error creating thumbnail: File missing

Abbildung 5.3: Vereinfachte Darstellung einer PCA auf nur zwei Dimensionen. Bei gleichzeitiger Betrachtung alle (zwei) Dimensionen sind hier deutlich zwei unterscheibare Gruppen zu erkennen. Reduziert auf eine einzige Dimension, X oder Y, zeigt sich in den Daten aber keine bimodale Verteilung; die Gruppen lassen sich nicht mehr unterscheiden. Ebenso kann es in einem Datensatz mit 100 oder mehr Dimensionen schwierig werden, jene Dimensionen (oder Kombinationen von Dimensionen) auszumachen, in denen Unterschiede deutlich werden. Die Achsen der beiden Principal Components, die sich für diesen Datensatz berechnen lassen, sind hingegen an die Varianzverteilung der Datenpunkte angepasst. Aus DARIAH-DE Report 5.2.3: Stand der Forschung in der Textanalyse.

Die PCA wurde erstmals von Karl Pearson[6] und Harold Hotelling[7] beschrieben. Sie erlaubt es, in einem hochdimensionalen Datensatz eine Betrachtungsebe der sich möglichst viel von der Varianz der Daten visuell erfassen lässt.

Error creating thumbnail: File missing

Abbildung 5.4: Entlang der neu berechneten Achse PC1 verläuft die Dichtekurve bimodal. Nun wird der Unterschied zwischen den beiden Gruppen schon in einer einzigen Dimension sichtbar. Aus DARIAH-DE Report 5.2.3: Stand der Forschung in der Textanalyse.

Hierfür werden die Dimensionen der Daten mit Hilfe der sog. Singulärwertzerlegung in ein neues Set von Variablen, die *Principal Components* (PC), transformi Principal Components kann man als Achsen eines alternativen Koordinatensystems verstehen, in dem die selben Datenpunkte in der selben Anordnung aufgetra

Achse dieses neuen Bezugsystems (PC1) führt exakt durch die Datenpunkte in Richtung ihrer größten Ausdehnung, sie beschreibt also die größte Varianz der Da Achsen (PC2 bis PCn) repräsentieren andere neue, orthogonal zur PC1 verlaufenden Achsen in Reihenfolge der Varianz, die der Datensatz in diesen Dimensione (Abb. 5.3). Folglich kann diese Technik eingesetzt werden, um aus einem Datensatz mit beliebig vielen Dimensionen eine zweidimensionale Darstellung (mit PC X- bzw. Y-Achse) zu erzeugen, die exakt diejenige Betrachtungsebene zeigt, in der der größte Teil der Datenvarianz zu sehen ist und oftmals auch die Unterschie Gruppen von Punkten am besten herausgestellt werden (Abb. 5.4).

Dieses rechnerisch aufwendige Verfahren fand mit Aufkommen des Computers zunehmend mehr Berücksichtigung in unterschiedlichen Bereichen wie beispiels Biologie, der Meterologie oder bei Bildkompressionsverfahren.

Im Bereich der Textanalyse setzten Mosteller und Wallace[8] die Methode zur Untersuchung der Federalist Papers erstmals im Zusammenhang mit Autorschafts Die PCA erlaubt hier, bei einer Vielzahl von Dimensionen, in denen man Unterschiede zwischen Gruppen vermutet, diejenige Betrachtunsgebene zu finden, in d Unterschiede am besten sichtbar werden. Vor allem, wenn es um die Zuordnung eines einzelnen Textes unbekannter Herkunft zu einem von zwei Autoren geht, mehrere sicher zugeordnete Vergleichstexte vorliegen, ist die PCA oftmals gut geeignet, die stilistische Ähnlichkeit zu einer der beiden Textgruppen visuell hera [11]. Aber auch zur Analyse der zeitlichen Entwicklung von Schreibstilen[12], oder der stilistischen Unterschiede zwischen Dialogen und narrativen Textpassagen PCA eingesetzt werden.

Die Messung stilistischer Distanzen

Error creating thumbnail: File missing

Abbildung 5.5: Der Abstand zweier Punkte A und B in einem Koordinatensystem: Manhattan-, Euklidische und Cosinus-Distanz. Aus Jannidis et al. 2015.

Noch weiter lässt sich die Analyse stilistischer Unterschiede operationalisieren, indem man diese auch tatsächlich quantifiziert. Die Modellierung von Texten als einem hochdimensionalen Koordinatensystem bietet hierbei die Möglichkeit, Abstände zwischen diesen Punkten direkt zu berechnen und als Maß für die stilisti Verschiedenheit zweier Texte zu verwenden. Es gibt in der Mathematik eine Reihe von Möglichkeiten, den Abstand zwischen zwei Punkten in einem mehrdimen zu messen. Drei davon kommen in stilometrischen Verfahren zum Einsatz: die ***Manhattan***-Distanz, d.h. die Summe aller Abstände in den einzelnen Dimensione **Euklidische** Distanz, d.h. die Länge der direkten Verbindungslinie zwischen den Punkten durch alle Dimensionen, und die **Cosinus**-Ähnlichkeit. Letztere fasst im Modell durch Reihen von Zahlenwerten repräsentiert werden nicht als Punkte auf, sondern als Vektoren, und quantifizieren deren Unterschiedlichkeit bzw. Ä Cosinuswert des Winkels zwischen den beiden Vektoren (Abb. 5.5).

Error creating thumbnail: File missing

Abb. 5.6: Texte zweier verschiedener AutorenInnen in einem vereinfachten, zweidimensionalen Feature-Raum. Die Texte der einen Autorin oder des einen Autors werden durch Kreise, die der/des anderen durch Dreiecke repräsentiert. Die stilistischen Abstände zwischen den Texten lassen sich in diesem Modell als Linien darstellen. Blaue Linien zeigen dabei Abstände zwischen Texten aus der gleichen Feder, rote Linien Vergleiche zwischen Texten unterschiedlicher Urheberschaft. Aus Jannidis et al. 2015.

Das erste Verfahren dieser Art, das in der Textanalyse erfolgreich war und bis heute in vielen Bereichen eingesetzt wird, wurde von John Burrows[14] vorgestellt **Burrows' Delta** bekannt gewordenen Verfahren werden die Worthäufigkeiten zunächst in relative Wortfrequenzen, d.h. in Prozent der Gesamtsumme aller Wört umgerechnet. Anschließend erfolgt eine sog. z-Transformation, die dafür sorgt, daß alle Werte mit einer Standardabweichung von Eins um einen Mittelwert von Ohne diese Standardisierung wäre das Gewicht der häufigsten Worte, wie "und", "der" und "die", so groß, daß die anderen Worthäufigkeiten gar keinen Einfluß Analyse haben, durch die Standardisierung haben alle Features vergleichbar große Werte und fallen gleichermaßen ins Gewicht. Auf den standardisierten relativ den sog. z-Scores, wird nun die Manhattan-Distanz berechnet. Dieser Wert wird als **Delta** bezeichnet, und dient als Maß für die Unterschiedlichkeit zweier Text sein Verfahren an einem Korpus mit Texten von 25 englischen Autoren aus dem 17. Jahrhundert. Es konnte dabei zeigen, daß sich ein Textabschnitt von nur 200 anhand von Delta-Abständen mit einer Erfolgsquote von 95% dem richtigen Autor zuordnen lässt, und das auf Basis von nicht mehr als den Häufigkeiten der 15 Wörter (Abb. 5.6).

Wenngleich John Burrows ursprüngliche Variante von Delta nach wie vor erfolgreich in der Forschung eingesetzt wird existieren mittlerweile mehrere Weiteren Argamon[15] schlug auf Grundlage mathematischer Argumente eine Variante vor, die statt der Manhattan-Distanz die Euklidische Distanz verwendet. Empirisch allerdings nicht zeigen, dass **Argamons Delta** in der Praxis bei der Autorenschaftszuschreibung besser funktioniert als Burrows Delta[16]. Rybicki und Eder[17] Variante, die speziell an die Bedürfnisse stark flektierter Sprachen wie Polnisch und Latein angepasst ist. Im Vergleich zu einer weitgehend unflektierten Sprach Englischen, ist bei Sprachen mit größerer morphologischer Formenvielfalt zu erwarten, daß die relative Häufigkeit der häufigen Wörter insgesamt weniger groß **Eders Delta** werden die *Features* nach ihrem Rang in der Liste der häufigsten Wörter gewichtet, um diesen Unterschied zu kompensieren. Die bisher beste Erfo empirischen Vergleich erreichte eine von Smith and Adrigde[18] vorgeschlagene Variante, bei der die Cosinus-Ähnlichkeit der z-Scores berechnet wird. Vor aller **Delta** auch bei sehr vielen *Features* stabil gute Ergebnisse, während die Erfolgsquote der anderen Varianten sinkt, wenn mehr als die 2000 häufigsten Wörter in eingehen[19]. Ein wesentlicher Grund dafür liegt vermutlich darin, dass in diesem Bereich der Wortliste zunehmend Worte auftreten, die nur in einzelnen Texten Frequenz vorkommen. Solche text-, und nicht autorenspezifischen Vokabeln können die Abstände zwischen Texten, die vom der gleichen Autorin/vom gleichen bei anderen Delta-Verfahren sehr groß werden lassen. Sie haben aber einen geringeren Effekt auf die Cosinus-Distanz, da die Wirkung einzelner Extremwerte hi Weise gedämpft wird wie nach einer **Vektor-Normalisierung** [20].

Stilometrische Analysen in *Stylo*

Für solche stilometrischen Analyseverfahren stehen heutzutage verschiedene, frei verfügbare Werkzeuge zur Verfügung. Eine der umfangreichsten Implementier stilometrischer Methoden bietet das ***Stylo*** -Paket von Maciej Eder, Jan Rybicki und Mike Kestemont. Es handelt sich dabei zwar im Prinzip um ein Packet für di R, erfordert aber keinerlei Programmierkenntnisse: Der Anwender kann über die R-Konsole eine graphische Benutzeroberfläche (*Graphical User Interface* oder über die sich die meisten Funktionen von Stylo per Mausklick bedienen lassen. Zur methodischen Grundausstattung von Stylo gehören sowohl die PCA, als auc von Texten anhand von Delta-Abständen.

Vorbereitung

Stylo zu nutzen erfordert zunächst einmal eine Installation von R. Aktuelle Installationsanleitungen für die gängigen Betriebssysteme finden sich auf der Projekt

https://www.r-project.org/

Nach der Installation kann R nun, entweder über die Programmverknüpfung, oder, in einem Unix-basierten Betriebssystem, über die Eingabe des Befehls "`R`" in Kommandozeile, gestartet werden. Innerhalb der **R-Konsole** sollte nun das **Paket "`stylo`" installiert** werden. Nutzt man R in einer graphischen Benutzeroberfl der Windowsversion automatisch mit installiert wird, so kann man Pakete aus dem zentralen CRAN-Repository normalerweise über das Menü installieren. Eine Möglichkeit, die unabhängig von Nutzeroberfläche und Betriebsystem überall gleich funktioniert besteht darin, in die R-Konsole den Befehl

```
install.packages("stylo")
```

einzugeben, und die Eingabetaste zu drücken. Dieser Befehl installiert das Paket, das nun mit einem weiteren Befehl

```
library(stylo)
```

geladen, d.h. aktiviert werden kann. (Auch hier muss nach dem Befehl die Eingabetaste betätigt werden.) **Wichtig:** Dieser Befehl ist auch dann nötig, wenn die Paketes über das Menü vorgenommen wurde, und muss bei jedem Neustart von R wiederholt werden.

Der nächste Schritt ist nun die Vorbereitung der zu installierenden Texte. Stylo nimmt sich die Texte für seine Analyse aus einem Unterverzeichnis namens "corp Arbeitsverzeichnisses. Zunächst einmal muss also auf dem Computer ein **Arbeitsverzeichnis angelegt** werden. Dieses könnte unter Windows beispielsweise "c: oder in einem Unixsystem "/home/MeineAnalyse/" oder "~/MeineAnalyse" heißen. In diesem Verzeichnis muss nun ein Unterordner namens "corpus" angelegt dorthin lautet dann also "c:\MeineAnalyse\corpus\"), in dem dann die **Texte abgelegt** werden. Für die Arbeit mit Stylo wird jeder Text in einer eigenen Datei ge als Formate sowohl TXT und HTML als auch TEI-XML in Frage kommen. Interessant ist hierbei insbesondere die **Benennung der Dateien**. Stylo verwendet di später als Beschriftungen in den Visualisierungen. Der erste Teil des Dateinamens, sofern mit einem Unterstrich abgetrennt, wird dabei als Gruppierungsvariable bildet die Grundlage für farbliche Unterscheidungen. Zur Untersuchung von Autorenschaftsfragen eignet sich also besonders ein Benennungsschema, das mit ei Autorennamen beginnt, der durch einen Unterstrich von einem eindeutigen Titel getrennt ist. Ein geeigneter Dateiname für Rudyard Kiplings "The Jungle Book "Kipling_TheJungleBook.txt". Die folgenden Code- und Analysebeispiele beziehen sich auf kleines Beispielkorpus von 12 englischsprachigen Kurzgeschichten verschiedenen Autoren, die alle in einem Zeitraum von etwa 50 Jahren entstanden sind.

Sobald nun die Ordnerstruktur steht, begibt man sich in die R-Konsole, um dort den gewählten Ordner (MeineAnalyse) als **Arbeitsverzeichnis** einzustellen. De lautet "setwd()" (das steht für set working directory) und könnte bei uns, je nach Betriebssystem (s.o.), Beispielsweise so aussehen:

```
setwd("~/MeineAnalyse/")
```

Wenn man sich nicht sicher ist, ob man schon im richtigen Arbeitsverzeichnis ist, kann das aktuelle Arbeitsverzeichnis in der R-Konsole auch mit dem Befehl

```
getwd()
```

abgefragt werden. Um Stylo nun zu **starten** und über das paketeigene GUI zu bedienen, gibt man als letzten Konsolenbefehl

```
stylo()
```

ein und drückt die Eingabetaste.

Nutzung

Hat man bis hierhin alles richtig gemacht, so sollte man nun das Fenster des Stylo-GUI vor sich sehen (Abb. 5.6).

Error creating thumbnail: File missing

Abbildung 5.6: Stylo GUI

In diesem Fenster können nun eine Reihe von Einstellungen vorgenommen werden. Das in unserem Beispiel verwendete TXT-Dateiformat entspricht bereits der Ebenso wie die Sprache (in diesem Beispiel Englisch) und die Verwendung einfacher Worthäufigkeiten als *Features*. An all diesen Einstellungen müssen für uns Veränderungen vorgenommen werden. Wählt man nun unter "STATISTICS" die "PCA (corr.) als Analyseverfahren und klickt auf "OK", so erzeugt Stylo die Vis ersten beiden Principal Components (Abb. 5.7).

Error creating thumbnail: File missing

Abbildung 5.7: PCA in Stylo

Hier zeigt sich bereits sehr deutlich, wie sich die Texte der vier Autoren in vier Gruppen aufteilen. Hätten wir zuvor eine der Dateien mit dem Autorennamen "U versehen, so könnten wir den Text nun trotzdem aufgrund seiner Position im Koordinatensystem einer Gruppe zuordnen. Zu beachten ist bei dieser Darstellung, größeren Teil der Datenvarianz repräsentiert (30,6% im vergleich zu 22,9% bei PC2), und dass eben nur zwei von 100 Dimensionen dargestellt werden, wenn au der größten Varianz. Folglich sollten vergleichende und quantifizierende Aussagen über die größere oder kleinere Ähnlichkeit zweier Autoren rein auf Basis der werden.

Für solche Aussagen bietet sich eher eine auf Delta-Abständen basierende Clusteranalyse an. Wählt man als Methode unter "STATISTICS" "Cluster Analysis" u "DISTANCES" "Classic Delta", dann erzeugt Stylo ein Baumdiagramm, bei dem die Entfernung zwischen den Texten entlang der Äste des Diagramms ihrer stil Unterschiedlichkeit nach Burrows Delta entspricht (Abb. 5.8).

Hier zeigt sich wieder klar eine Gruppierung der Texte nach ihren Autoren. Gleichzeitig aber spaltet sich das Baumdiagramm schon früh in zwei Untergruppen; britischen Autoren Doyle und Kipling finden sich auf dem einen Ast, was zeigt, dass sie sich sich stilistisch besonders ähnlich sind, die beiden Amerikaner Love auf dem anderen.

NLP-Tools in der Stilometrie

Was ist nun aber, wenn man sich für andere Features interessiert, wenn das Abzählen der häufigsten Wörter nicht ausreicht, oder von vornherein ungeeignet ersc Forschungsfrage zu beantworten? Was, wenn man eher das Inventar an beschreibendem Vokabular vergleichen möchte, oder den Satzbau? Stylo selbst bietet die Wörter auch Buchstaben oder Zeichen als Features zu verwenden, oder sog. n-Gramme, als Ketten von Worten oder Zeichen in einder definierbaren Länge. Will tatsächlich an bestimmte Wortklassen oder Satzstrukturen heran, wird der Einsatz zusätzlicher Werkzeuge aus dem Bereich des *Natural Language Processing* (N

ComputerlinguistInnen haben in den vergangen Jahren eine ganze Reihe solcher Werkzeuge entwickelt und arbeiten stetig an ihrer Verbesserung. So können heu Reihe von Sprachen diverse linguistische Analysen automatisiert durchgeführt werden, dazu gehören z.B. die Lemmatisierung, Grammatische Satzanalyse und d Eigennamen. Das **DKPro-Projekt** der Technischen Universität Darmstadt entwickelt eine Programmierumgebung, in der viele dieser unabhängig voneinander e Werkzeuge zu einer Art virtueller Fließbandverarbeitung zusammen gebaut werden können, um komplexe, mehrstufige linguistische Analyseprozesse zu realisie Verwendung von DKPro erfordert allerdings Programmierkenntnisse in Java, was grundsätz eine recht hohe Einstiegshürde darstellt. Um auch DH-affinen GeisteswissenschaftlerInnen ohne diese Kenntnisse den Zugang zu den Möglichkeiten von DKPro zu bieten, wurde im Rahmen von DARIAH-DE der sog. **DK** entwickelt, ein fertig zusammegebautes DKPro-Programm, das sich als Java-Datei herunterladen, in der Kommandozeile ausführen, und über eine Konfiguratio lässt. Der DKPro-Wrapper erzeugt aus einer Textdatei eine **CSV-Tabelle**, in der der ursprüngliche Text Wort für Wort in einer Spalte steht, und in den anderen S diversen, computergenerierten linguistischen Annotationen versehen ist. Dieses Output-Format ist für die Weiterverarbeitung in Datenanalyseskripten, wie man [www.r-project.org|R] oder mit Hilfe des Pandas-Paketes in Python schreiben kann, es lässt sich aber grundsätzlich auch einfach als Tabelle in MS-Excel öffnen

Ein ausführliches Tutorial, das beschreibt, wie der DKPro-Wrapper installiert und ausgeführt wird, wie sich die Konfiguration an eigene Bedürfnisse anpassen lä mit dem Output sowohl stilistische, als auch inhaltliche Analysen durfüren kann, findet sich im von DARIAH-DE. Hier nur die kurze Version. Zunächst einma eine aktuelle Installation des SE Development Kit. Die Aktuelle Version des DKPro-Wrappers kann bei [1] herunter geladen werden. Sie wird in das gewünschte Arbeitsverzeichnis kopiert und dort von der Kommandozeile aus, d.h. in der Unix-Shell oder über die Windows-Eingabeaufforderung, mit folgendem Befehl aus

```
java -Xmx4g -jar DateinameVomAktuellenDKProWrapper.jar -input PfadZurTextdatei.txt -output PfadZumOutputordner
```

Die Option -Xmx4g ist notwendig, um Java ausreichen Arbeitsspeicher für die Rechenoperation zuzuweisen. Konkret könnte der Befehl also so aussehen:

```
java -Xmx4g -jar de.tudarmstadt.ukp.dariah.pipeline-0.3.0-standalone.jar -input C:\MeineAnalyse\corpus\Kipling_TheJungleBook.txt -output C:\M
```

Error creating thumbnail: File missing

Abbildung 5.8: Baumdiagramm einer Clusteranalyse
basierend auf Delta-Abständen
Error creating thumbnail: File missing

Abbildung 5.9: Output des DKPro-Wrapper geöffnet in
LibreOffice Calc.

Wichtig: Das Programm ist darauf angewiesen, temporär Komponenten aus dem Internet nachzuladen, es erfordert also eine funktionierende Internetverbindung **wichtig:** Je nach Hardware, Länge der Textdatei und Auswahl an Analysverfahren, die der DKPro-Wrapper durchführen soll, kann alles **sehr lange dauern**!

Wenn der Wrapper durchgelaufen ist, sollte er eine CSV-Datei mit den Analyseergebnissen erzeugt haben. Diese Ergebnisse können nun wiederum in Stylo eing Prinzipiell kann daraus ein Textkorpus für Stylo von Hand erzeugt werden, indem man die relevante Spalte aus dem CSV kopiert und in einer neuen Textdatei m geeigneten Benennung in einem Ordner namens "corpus" speichert. Eleganter lässt sich das natürlich mit einem kurzem Skript in R oder Python erledigen, hier l zusätzlich die Möglichkeit, auch bedingte Abfragen unter Einbeziehung mehrerer Spalten zu implementieren um beliebig komplexe Features zu extrahieren. Das Skript (das man auch kopieren und in der eigenen R-Konsole ausführen kann) zeigt beispielhaft, wie wie man nach der Verarbeitung im DKProWrapper aus den werwendeten Korpus zwei verschiedene neue Korpora für die Analyse in Stylo erstellt. Zum einen wird ein unterordner erstellt, in dem alle Texte auf lemmatisie Adverbien, also auf ihr Deskriptives Vokapular ("dv") reduziert sind. Im anderen Ordner ("pos") finden sich in den Textdateien statt der urspünglichen Sätze nur entsprechenden grammatikalischen Funktionsbezeichnungen (engl. "*part-of-speech* tags" oder "POS-tags").

```
# Extract file names
files = list.files(pattern = "*.csv")

# Create directories
dir.create("dv/")
dir.create("pos/")
dir.create("dv/corpus/")
dir.create("pos/corpus/")

for(file in files){
 # Read file
 df = read.table(file, header = T, fill = T)

 # Prepare filename
 shortfile = sub(".csv", "", file)

 # Write lemmatized Adjectives and Adverbes to analyse the author's inventar of descriptive vocabulary
 dv = df$Lemma[df$CPOS == "ADJ" | df$CPOS == "ADV"]
 filename = paste("./dv/corpus/", shortfile, sep = "")
 write(paste(dv, collapse = " "), file = filename)

 # Write POS tags to compare sentence structure
 filename = paste("./pos/corpus/", shortfile, sep = "")
 write(paste(df$CPOS, collapse=" "), file = filename)
 }
```

In den Ordnern "dv" und "pos" befindet sich nun jeweils wieder ein Ordner mit dem Namen "corpus", in dem die Dateien abgelegt sind. Es genügt also, in gewo R-Konsole einen der neuen Unterordner als Arbeitsverzeichnis auszuwählen und Stylo zu starten:

```
setwd("./dv/")
stylo()
```

Sollte hier die Fehlermeldung

```
Error: could not find function "stylo"
```

erscheinen, so hat man vermutlich vergessen, das Paket vorher zu laden. Zur Erinnerung: `library(stylo)`.

In der Nutzeroberfläche kann nun, wie oben, die Clusteranalyse auf Delta-Basis ausgewählt werden. Bei Betrachtetung des nun erzeugten Baumdiagramm für da Vokabular (Abb. 5.10) zeigt sich zunächst, dass die Texte auch bei dieser Analyse zunächst nach Autoren gruppiert werden. Allerdings teilt sich der Baum nun n britischen und amerikanischen Autoren auf, dafür zeigen sich deutliche Ähnlichkeiten zwischen Howard Kipling. (Man mag nun spekulieren, ob das mit der Vor beiden Autoren für Geschichten in exotischen, oft tropischen Umgebungen zusammen hängt.)

Error creating thumbnail: File missing

Abbildung 5.10: Baumdiagramm der Stilistischen Ähnlichkeiten basierend auf dem deskriptiven Vokabular der Autoren.

Für die Analyse der Satzstruktur wechselt man mit dem Arbeitsverzeichnis in den beachbarten Ordner und startet Stylo dort:

```
setwd("../pos/")
stylo()
```

In den Textdateien befinden sich in diesem Fall nur die Part-of-Speech Tags. Unter "FEATURES" wählt man in Stylo nun die "N-GRAM SIZE" von 3, um Drei Tag als Features in die Analyse einzuspeisen. Individuelle Vorlieben beim Satzbau sollten sich also durch charakteristische Dreierkombinationen in der Analyse Bei dieser Analyse (Abb. 5.11) verhalten sich Kiplings Texte auffällig anders als die der anderen Autoren. Während letztere nach wie vor dicht zusammen liegen den Satzbau bezogen eine starke stilisitische Variabilität innerhalb von Kiplings verschiedenen Geschichten.

Error creating thumbnail: File missing

Abbildung 5.11: Baumdiagramm der Stilistischen Ähnlichkeiten basierend auf dem Satzbau.

Links und Literatur

- Verfahren der Digital Humanities in den Geistes- und Kulturwissenschaften". *DARIAH-DE Working Papers* Nr. 4. Göttingen: DARIAH-DE, 2014. URN: dariah-2014-2-6
- DARIAH-DE Geo-Browser Dokumentation https://dev2.dariah.eu/wiki/display/publicde/Geo-Browser+Dokumentation
- DARIAH-DE Datasheet Editor Dokumentation https://dev2.dariah.eu/wiki/display/publicde/Datasheet+Editor+Dokumentation
- Matthew L. Jockers, Macroanalysis: Digital Methods and Literary History (University of Illinois Press, 2013)
- Thomas Kollatz; Stefan Schmunk: Datenvisualisierung: Geo-Browser und DigiVoy. In: TextGrid: Von der Community für die Community – Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften. 2015, S. 165-180 (http://www.univerlag.uni-goettingen.de/handle/3/Neuroth_TextGrid).
- Franco Moretti: Distant Reading, 2013
- Stefan Pernes und Steffen Pielström, 2015: Die quantitative Analyse großer Datenbestände in den Geisteswissenschaften: eine Kommentierte Bibliograph Report 5.2.2
- S. Bock, K. Du, P. Dürholt, T. Gradl, M. Huber, M. Munson., S. Pernes und S. Pielström, 2015: Stand der Forschung in der Textanalyse. DARIAH-DE Re

Anmerkungen

1. ↑ Tools und Dienste in DARIAH-DE https://de.dariah.eu/tools-und-dienste
2. ↑ http://dirtdirectory.org "The DiRT Directory is a registry of digital research tools for scholarly use. DiRT makes it easy for digital humanists and others c research to find and compare resources ranging from content management systems to music OCR, statistical analysis packages to mindmapping software.
3. ↑ https://de.dariah.eu/geobrowser
4. ↑ http://geobrowser.de.dariah.eu
5. ↑ http://geobrowser.de.dariah.eu/edit/
6. ↑ Karl Pearson. “On lines and planes of closest fit to systems of points in space”, Philosophical Magazine, Series 6, vol. 2, no. 11, 1901, pp. 559-572.
7. ↑ Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417–441, and 498

8. ↑ Frederick Mosteller and David L. Wallace. 1964. Inference and Disputed Authorship: The Federalist. SpringerVerlag, New York. 2nd Edition appeared i called Applied Bayesian and Classical Inference.
9. ↑ Burrows J, 1989. " 'An ocean where each kind...': statistical analysis and major determinants of literary style". Computers and the Humanities 23: 309-3
10. ↑ Binongo und Smith, "The Application of Principal Component Analysis to Stylometry", Literary and Linguistic Computing 14.4, 1999.
11. ↑ Binongo JNG, 2003. "Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution". Chance 16(2): 9-17
12. ↑ Brainerd B, 1980. "The chronology of Shakespeare's plays: a statistical study". Computers and the Humanities 14: 221-230
13. ↑ Burrows J, 1987. "Word patterns and story-shapes: the statistical analysis of narrativve style". Literary and Linguistic Computing 2(2): 61-70
14. ↑ John Burrows: "Delta: A Measure for Stylistic Difference and A Guide to Likely Authorship". In: LLC 17,3 2002.267-87.
15. ↑ Shlomo Argamon: "Interpreting Burrows's Delta: geometric and probabilistic foundations". Literary and Linguistic Computing 2008;23(2):131-47.
16. ↑ Fotis Jannidis, Steffen Pielström, Christof Schöch and Thorsten Vitt. 2015. "Improving Burrows' Delta - An empirical evaluation of text distance measu Humanities Conference 2015.
17. ↑ Jan Rybicki and Maciej Eder: "Deeper Delta across genres and languages: do we really need the most frequent words?" Lit Linguist Computing (2011)
18. ↑ Smith, Peter WH, and W. Aldridge. "Improving Authorship Attribution: Optimizing Burrows' Delta Method*." Journal of Quantitative Linguistics 18.1 (
19. ↑ Fotis Jannidis, Steffen Pielström, Christof Schöch and Thorsten Vitt. 2015. "Improving Burrows' Delta - An empirical evaluation of text distance measu Humanities Conference 2015.
20. ↑ Evert S, Proisl T, Jannidis F, Pielström S, Schöch C und Vitt T, 2015. "Towards a better understanding of Burrows's Delta in literary authorship attributi HLT Fourth Workshop on Computational Linguistics for Literature, Denver, Colorado.

Forschungsinfrastrukturen nutzen

Ziele und Grundlagen einer Forschungsinfrastruktur

Werkzeuge und Methoden, die wir aus den angewandten Wissenschaften und den Naturwissenschaften kennen, lassen sich nicht ohne Weiteres auf die Geistes- Kulturwissenschaften übertragen. Eine Forschungsinfrastruktur selbst kann aber als Werkzeug für die digitale geistes- und kulturwissenschaftliche Forschung ve Dieses Werkzeug besteht aus vielen verschiedenen Komponenten und erlangt durch bedarfsgerechte Vielfalt, die in Zusammenarbeit mit den Fachwissenschaftle und entwickelt wird, seine Mächtigkeit. Es ist jedoch darauf zu achten, die Balance zwischen sehr fachspezifischen Anforderungen einerseits und möglichst gen Entwicklungen für alle Disziplinen andererseits zu halten. Nur so kann es gelingen, komplexe Forschungsfragen aus unterschiedlichsten und fachlich übergreife Blickwinkeln formulieren zukönnen.

Forschungsinfrastrukturen (FI) sind kein statisches Produkt, sondern sind als kontinuierlicher Prozess zu betrachten. Dieser Prozess beinhaltet auch, wie verschi Forschungsverbünde zeigen, eine starke Vernetzung der FachwissenschaftlerInnen. Nur durch den Zusammenschluss von WissenschaftlerInnen in nationalen un Kooperationsvorhaben und Kollaborationen können wichtigen Forschungsfragen beantwortet und Erkenntnisse erzielt werden.

Im Zeitalter von e-Science bzw. e-Research werden Theorie, Experiment und Simulation zusammengeführt und durch eine Forschungsinfrastruktur unterstützt. Forschungsinfrastruktur selbst besteht dabei aus vielfältigen Angeboten und Diensten in den Kernbereichen **Forschung**, **Lehre**, **Forschungsdaten** sowie **Techn Infrastruktur**[1]. Lediglich im Zusammenspiel aller Bereiche lassen sich digital arbeitende WissenschaftlerInnen optimal fördern.

Forschung: Zur Etablierung einer digitalen Forschungskultur ist das Wissen um entsprechende Forschungsmethoden und Verfahren in den Geisteswissenschafte Einsatz dieser Methoden und Verfahren wird durch digitale Dienste und Werkzeuge unterstützt, die konzipiert, entwickelt und als Basisinfrastruktur bereitgestell konkreter Forschungsfragen werden zunächst die fachwissenschaftlichen Bedürfnisse identifiziert. Basierend darauf können einzelne Lösungen als sogenannte D bzw. fachwissenschaftliche Dienste entwickelt werden, um exemplarisch Potentiale, Methoden und Leistungen der Digital Humanities aufzuzeigen.

Lehre: Der kompetente Umgang mit digitalen Ressourcen, Konzepten und Methoden der Digital Humanities muss Eingang in die Lehre und Fortbildung von GeisteswissenschaftlerInnen auf allen Stufen ihrer Ausbildung und beruflichen Praxis finden. Eng vernetzt mit den fachwissenschaftlichen Communities werden Studien- und Weiterbildungsangebote abgestimmt, besser sichtbar gemacht und weiterentwickelt. Darüber hinaus können eigene Qualifizierungsmodule, wie z.B fokussierte, internationale Expertenworkshops, angeboten werden.

Forschungsdaten: Forschungsdaten spielen eine zentrale Rolle im gesamten Forschungsprozess von der Recherche und Erfassung über die Analyse und Verarb Publikation und anschließenden Nutzung auch durch Dritte. Eine wesentliche Grundlage dafür ist der vertrauensvolle und ungehinderte Zugriff auf diese Forsch ergebnisse. Fachrelevante Standards für Daten, Metadaten, Lizenzen, Werkzeuge sowie Prozeduren und Organisationsstrukturen müssen evaluiert und diskutiert Empfehlungen abgeben zu können.

Technische Infrastruktur: In diesem Bereich werden Software-, Plattform- und Infrastruktur-Hosting-Services sowie operative IT-Dienste als tragfähige und fl Lösungen entwickelt und bereitgestellt. Beispiele für Komponenten einer technischen Infrastruktur finden sich im nächsten Kapitel.

Aufbau einer Forschungsinfrastruktur am Beispiel von DARIAH-DE

Wer digital forscht und Verfahren der Digital Humanities anwendet, benötigt entsprechende Werkzeuge und Dienste. Operative IT-Dienste bilden die Grundlage Entwicklungstätigkeiten und den Betrieb des technischen Teils einer Forschungsinfrastruktur. Zu diesem Bereich zählen beispielsweise

- eine **kollaborative Arbeitsumgebun**g zur gemeinsamen Bearbeitung von Texten und Code,
- die **Bereitstellung virtueller Maschinen** für eine Vielzahl von Diensten,
- eine nachhaltige **Speicherung von Forschungsdaten**,
- eine Authentifizierungs- und Autorisierungsinfrastruktur zur Sicherungstellung eines **sicheren Zugriffs** und einer sicheren Nutzung von Angeboten und F
- ein zentrales **Monitoring von Diensten und Werkzeugen** und
- eine zentralen **Anlaufstelle für alle Fragestellungen** rund um die Infrastruktur.

Auf dieser Basis fußt der Aufbau von 'höherwertigen' und fachwissenschaftlichen Diensten und Angeboten einer digitalen Forschungsinfrastruktur für die Kultu Geisteswissenschaften. In diesem Abschnitt werden exemplarisch Komponenten der Forschungsinfrastruktur von *Digital Research Infrastructure for the Arts an* (DARIAH-DE) beschrieben, die für fachwissenschaftliche Dienste bereits genutzt werden. Diese und weitere Informationen finden sich hier[2].

Kollaborative Arbeitsumgebung

Der wachsende Bedarf nach kollaborativen digitalen Methoden und Werkzeugen wird von DARIAH-DE aufgegriffen und im Rahmen des Aufbaus und Betriebs Forschungsinfrastruktur realisiert. Durch kollaborative Arbeitsumgebungen können digitale Werkzeuge und Dienste – unabhängig von Betriebssystem, Software Standort – gemeinsam genutzt und Daten und Texte zeitgleich bearbeitet werden. Das gemeinsame Arbeiten an einem digitalen Projekt wird durch ein differenzi Rechtemanagement (Rechte- und Rollenzuweisung) wesentlich erleichtert und ermöglicht die Zusammenarbeit in einer geschützten Umgebung. Dank der Bandl

Werkzeugen und Forschungsdaten können neue Fragen an alte Forschungsgegenstände formuliert werden oder auch neue Forschungsfragen entstehen. Durch di Vernetzung von WissenschaftlerInnen können außerdem Forschungsthemen vertieft, ein aktiverer Austausch befördert und Forschungsergebnisse leichter von ar WissenschaftlerInnen nachgenutzt werden.

DARIAH-DE bietet verschiedene Dienste für das kollaborative Arbeiten am Forschungsgegenstand (z.B. Text-Quellen) an. Forschungsdaten und Quellen könne **DE Repository** oder im **TextGrid Repository**[3] gespeichert und von für das jeweilige Projekt autorisierten Personen orts- und zeitunabhängig mit Hilfe von ve Werkzeugen und Diensten bearbeitet werden. Im **Etherpad**[4] können mehrere Personen ortsunabhängig, aber zeitgleich gemeinsam einen Text erstellen und scl können Forschende ihre (Teil-)Ergebnisse sammeln, bearbeiten und ihre Dokumentation auch anderen Personen zur Verfügung stellen. Das angebotene Wiki-Sy Möglichkeit, einen passwortgeschützten internen und einen zugriffsfreien öffentlichen Bereich einzurichten.

Ergänzend dazu werden im **Developer-Portal**[5] weitere Werkzeuge und Softwarekomponenten für die Realisierung von DH-Projekten bereitgehalten. Das Dev basiert auf einer Reihe von Standard-Entwicklerwerkzeugen, die für Forschungs- und Entwicklungsprojekte der Digital Humanities "on demand" und flexibel ei können. Aktuell stehen im Developer Portal folgende Werkzeuge zur Verfügung, die auf vielfältige Weise Entwicklungsprozesse unterstützen:

- Confluence[6] - Dokumentation von Ergebnissen und Projektmanagement
- E-Mail-Liste[7] - einfache Erreichbarkeit aller Projektmitarbeiter
- SVN[8] - Versionskontrollsystem, das für gemeinsame Softwareentwicklung genutzt werden kann
- Jira[9] - Dokumentation und Management von Aufgaben (nicht nur für Software)
- Projektverwaltung[10] - einfache, online Verwaltung von Projekten und Issues
- Jenkins[11] - Unterstützung beim automatischen Kompillieren und Testen von Software

Verschiedene Nutzungsszenarien und viele Aufgaben können durch das vielfältige Angebot an Werkzeugen, das laufend entsprechend den Anforderungen und B DH-Developer-Community erweitert wird, abgedeckt werden.

Bereitstellung virtueller Maschinen

Die Bereitstellung von digitalen Ressourcen ist ein essentieller Baustein jeder verteilten Forschungsinfrastruktur. Das DARIAH-DE Hosting von Virtuellen Mas technisch vergleichbar mit den Angeboten von bekannten kommerziellen Cloud-Diensten, wobei die von DARIAH-DE für Geistes- und KulturwissenschaftlerIr VMs besonders auf die Anforderungen und Bedürfnisse dieser Disziplinen zugeschnitten sind.

Error creating thumbnail: File missing

Bild: Bereitstellung virtueller Maschinen, Quelle:
https://de.dariah.eu/documents/10180/356681/VMs_DARIAH.png

Die beteiligten Rechen- und Datenzentren können in diesem Zusammenhang auf langjährige Erfahrungen zurückgreifen. Die Bereitstellung von Virtuellen Masc DARIAH-DE:

- Zugriff auf VMs mit vorinstalliertem und konfiguriertem Betriebssystem
- VMs für Testzwecke und Produktionservices
- High-End-Ressourcen (Rechenleistung, Speicher, Netzwerk)
- Grundkonfiguration der Systeme (Firewall usw.)
- Einbindung in das DARIAH-DE Monitoring
- Sicherung der Dateisysteme

Bereitstellung von Speicher

Eine verlässliche, nachhaltige und persistente Speicherung von Daten ist die Grundvoraussetzung für jedes Forschungsprojekt. Die konkreten Anforderungen kö deutlich voneinander abweichen. Die Daten der Projekte und Forscher unterscheiden sich in ihrer:

- Größe (von einigen wenigen Kilobyte für einen Brief in einer Textdatei bis zu vielen Gigabyte für eine Filmaufnahme einer Oper),
- Menge (von einigen Bilddateien eines seltenen, wertvollen Manuskripts bis zu mehreren Millionen Bilddateien einer gesamten Bibliothek)
- und Typ, da es eine Vielzahl unterschiedlicher Formate für Text, Bild, Audio und Video gibt.

Um eine nachhaltige Referenzierung der gespeicherten Daten zu ermöglichen, wird die Verwendung von persistenten Identifikatoren (*Persistent Identifier* oder I Dienste wie das DARIAH-DE Repositorium erlauben eine umfassendere Preservation und die Kombination verschiedener Komponenten.

Nachhaltige Speicherung von Bitstreams

Error creating thumbnail: File missing

DARIAH Bitstream Preservation, Quelle:
https://de.dariah.eu/dariah-svg/small/130920-BitPreservation_690.png

Die DARIAH Bit Preservation wurde zur nachhaltigen, sicheren und persistenten Speicherung heterogener geisteswissenschaftlicher Forschungsdaten entwickel die folgenden Eigenschaften gekennzeichnet[12]:

- Daten werden unabhängig von Größe, Format oder Inhalt gespeichert.
- Nur administrative Metadaten werden erstellt und verwaltet. Inhaltlich erschließende Metadaten werden auf dieser Ebene als Datei behandelt.
- Es werden hauptsächlich CREATE und READ Operationen angewendet. Methoden zum Aktualisieren (UPDATE) oder Löschen (DELETE) sind verfügba allerdings selten bzw. nur administrativ genutzt.
- Es werden Mechanismen zur Sicherstellung der Datenintegrität bereit gestellt.
- Der Zugriff wird sowohl über intuitive, von Forschern einfach zu nutzende als auch über maschinenlesbare Schnittstellen ermöglicht.
- Durch Nutzung der DARIAH Authentifizierungs- und Autorisierungsinfrastrukture werden unerlaubte Zugriffe und Modifikationen verhindert.

Ein besonderer Fokus liegt auf Modularität und technologischer Nachhaltigkeit. Durch eine Speicherabstraktionsschicht können die anbietenden Institutionen zu vorhandenen Speichersysteme nutzen, zum anderen werden Datenmigrationen bei veralteter Software ermöglicht. Als Software wird in diesem Kontext zur Zeit eingesetzt, eine relativ weit bekannte *Open-Source Software* (OSS) zur redundanten Verwaltung und Speicherung großer Datenmengen.

Für die Interaktion mit der Bit Preservation wurden zwei Schnittstellen spezifiziert, die DARIAH Storage API und die DARIAH Admin API. Beiden Schnittstel den Standards HTTP- und REST, die sich im Bereich Webservices durchgesetzt haben. Die DARIAH Storage API bietet Funktionalitäten zur einfachen Speiche die DARIAH Admin API erlaubt Interaktion mit der Bitstream Preservation Komponente des Systems. Unter anderem können von den anbietenden Institution s Preservation Level festgelegt werden, die unterschiedlichen Güten der Preservationsmaßnahmen beschreiben. Zusätzlich können Informationen zu den Dateien Lokation der Repliken, verwendeter Prüfsummenalgorithmus, Häufigkeit der Integritätsüberprüfungen usw. abgefragt werden.

Nachhaltige Referenzierung von Digitalen Objekten

Error creating thumbnail: File missing

PID-Service in DARIAH-DE, Quelle:
https://de.dariah.eu/dariah-svg/small/PID_690.png

Kontinuierlich steigt die Menge der digital gespeicherten Daten in allen Bereichen der Forschung an. Die Verwaltung der Daten wird dadurch zunehmend komp insbesondere die nachhaltige Referenz auf Daten und ihre dauerhafte Zitierbarkeit eine große Herausforderung darstellt. Verweise können durch einen persistent (*Persistent Identifier*oder PID), der eine Mittlerrolle einnimmt, somit stabil bleiben, auch wenn sich der Speicherort der Daten ändert.

PIDs können in vielfältiger Art eingesetzt werden, beispielsweise um Daten und Metadaten zu digitalen Objekten zu bündeln, stabile Links für Publikationen be die Datenarchivierung zu organisieren. Voraussetzung ist eine entsprechende Pflege der Links, auf die die PIDs verweisen. Aus diesem Grund sollten persistente jeweils in Zusammenhang und in Absprache mit einer Institution (zum Beispiel einem Institut, einem Datenzentrum, einem Rechenzentrum oder einem Reposit dauerhafte Stabilität der PID-Verweise sorgen kann, verwendet werden.

Es existieren verschiedene PID Anbieter, deren Technologien ermöglichen, Identifikatoren zu erzeugen, zu verwalten und aufzulösen. Jeder PID Anbieter trifft b Grundannahmen über Daten und deren Verwendung. Der Ansatz der DARIAH Forschungsinfrastruktur sieht beispielsweise vor, dass PIDs bereits früh im Beart verwendet werden, selbst wenn noch nicht klar ist, ob das Objekt in eine Archivierung überführt wird und somit dauerhaft erhalten bleibt.

Die Anforderungen von DARIAH-DE stellen vergleichsweise wenig Grundbedingungen an die zu referenzierenden digitalen Objekte, was durch den PID Anbie werden muss. In DARIAH-DE wurde aus diesem Grund der EPIC PID Service als vertrauenswürdiger Persistenter Identifier Dienst gewählt. Das *European Per Consortium* (EPIC)[14] wurde 2009 mit dem Ziel gegründet, den europäischen Forschungsgemeinschaften einen einfachen PID-Service bereitzustellen. Unter de befinden sich Projekte wie EUDAT[15], CLARIN[16], TextGrid[17] und DARIAH. Über den EPIC Dienst ist es möglich, so genannte Handle[18] Identifikatoren verwalten und aufzulösen.

Der vergebene Identifier ist im Gegensatz zu anderen Ansätzen relativ flexibel in ihrer Gültigkeit, da im EPIC Standard der Gültigkeitszeitraum eines Identifiers werden kann. Dieses System kann zum Beispiel zur Behandlung von Objekten unterschiedlicher Granularität und deren flexiblen Kombination zu neuen Objekt

von besonderem Interesse im DARIAH Repository[19] sein. Zudem können EPIC PIDs, falls dies vom WissenschaftlerInnen gewünscht ist, problemlos in Digita Identifier (DOI)[20] PIDs überführt werden, da beide Anbieter eine Technologie auf Basis von Handle-Identifiern verwenden.

Sichere Dienste und Daten

Error creating thumbnail: File missing

DARIAH Authentifizierungs- und Autorisierungsinfrastruktur (AAI), Quelle: https://de.dariah.eu/dariah-svg/small/130904-AAI_690.png

Eine *Authentifizierungs- und Autorisierungsinfrastruktur* (AAI) ist für den Betrieb einer Forschungsinfrastruktur unabdingbar. Zum einen muss sichergestellt we Anfragen einer BenutzerIn wirklich von ihr stammen (**Authentifizierung**) und dass sie zu dieser Operation auf eine bestimmte Ressource berechtigt ist (**Autori** diesem Zweck können so genannte Attribute von einer zentralen Instanz abgefragt werden, beispielsweise Zugehörigkeit zu einer Nutzergruppe, und zur Entsche herangezogen werden.

Die Authentifizierungs- und Autorisierungskomponente von Forschungsinfrastrukturen basieren auf internationalen Standards, die sich im letzten Jahrzehnt sow Hochschulbereich als auch in der Industrie durchgesetzt haben. Auf diese Weise wird ein gemeinsames Vokabular verwendet, das die Authentifizierung der Benu teilnehmenden Einrichtungen bei den gewünschten Forschungsinfrastrukturen ermöglicht. In DARIAH-DE werden zur Zeit der *Security Assertion Markup Lang* Standard[21] sowie die open source Implementierung Shibboleth[22] verwendet.

Forschungsinfrastrukturen sind meist Teil eines Verbundes, so genannten Föderationen, um Ressourcen bereitzustellen und nutzen zu können. DARIAH-DE ist l Teil der deutschen Hochschulföderation DFN-AAI[23]. Auf diese Weise können sich auch MitarbeiterInnen von deutschen Hochschulen und Forschungseinrichtu DARIAH-DE Dienste authentifizieren. Im Rahmen des europäischen Geant-Projekts nimmt die DFN-AAI an der Meta-Föderation eduGain[24] teil, so dass die l BenutzerInnen der dort angebundenen nationalen Föderationen offen stehen. Eine zusätzliche Komponente der DARIAH-DE Benutzerverwaltung erlaubt ebenfa BenutzerInnen ohne Zugehörigkeit zu einer Forschungseinrichtung der angebundenen Föderationen die Teilhabe an Ressourcen, Projekten und Diensten einer Forschungsinfrastruktur.

Monitoring von Diensten

Error creating thumbnail: File missing

Monitoring in DARIAH-DE, Quelle: https://de.dariah.eu/dariah-svg/small/130905_Monitoring_690.png

Zum Betrieb der digitalen Forschungsinfrastrukturen gehört die Überwachung (das Monitoring) von Infrastrukturkomponenten und Diensten. Durch Monitoring auftretende Probleme identifiziert und Ausfälle schnellstmöglich behoben werden. Bei vielen Ressourcenanbietern sind bereits Systeme in Betrieb, die Server u überwachen. Aus der Sicht der Anbieter stehen allerdings vor allem die eigene Hardware und der Zustand der Basisdienste im Fokus.

Das Monitoringsystem einer Forschungsinfrastruktur soll sowohl EndnutzerInnen als auch AdministratorInnen gerecht werden. Neben der Verarbeitung der bere Monitoringinformationen der Anbieter muss es zusätzlich die bereitgestellten fachwissenschaftlichen Dienste integrieren, was auf Grund ihrer Komplexität in v triviale Aufgabe ist.

Das Monitoring ermöglicht die Überwachung entfernter Server und Dienste, die Sicherstellung der Erreichbarkeit der Systeme, sowie die Abfrage des allgemein Systemzustands. Durch die verwendete Software kann der Status der integrierten Systeme und Dienste visualisiert und die Einhaltung von Dienstgütevereinbaru werden. Im Falle eines Ausfall werden Verantwortliche automatisch benachrichtigt, um die Verfügbarkeit des Dienstes durch ein möglichst schnelles Eingreifen gewährleisten.

Zentrale Unterstützung bei Fragen

Beim Umgang mit einer Forschungsinfrastruktur können vielfältige Fragen entstehen, die mit den technischen und/oder fachwissenschaftlichen Experten bespro müssen. Eine sorgfältige Dokumentation der behandelten Themenfelder ist für alle ForscherInnen hilfreich, da viele Fragestellungen mehrmals auftreten und auf

direkt gelöst werden können. Ein so genanntes Support- oder Helpdesk-System unterstützt eine unkomplizierte und effektive Anfrage- sowie Supportbearbeitung Support- oder Helpdesk-System soll sichergestellt werden, dass keine Anfrage oder Nachricht verloren geht. Zusätzlich ermöglicht ein solches System jederzeit Gesamtüberblick über die zu bearbeitenden Vorgänge und kann den gesamten Verlauf einer Anfrage inklusive aller Antworten dokumentieren.

Die Anfragen, Aufgaben und Fragestellungen werden in sogenannten "Tickets" zusammengefasst und abgearbeitet. Ein Ticket ist prinzipiell ein Container für di zwischen den beteiligten Personen wie beispielsweise Fragesteller, Bearbeiter oder Manager bezüglich einer Anfrage. Die verwendete Software hilft beim Empf Bestätigung, bei der Klassifizierung und bei der Bearbeitung von Anfragen bzw. Tickets. Es besteht die Möglichkeit Tickets in verschiedene Bereiche wie beispi oder "Speicherplatz" einzuteilen und auf diese Weise direkt einer Person oder Personengruppe zur Bearbeitung und Lösung zuzuweisen. Bei jeder Aktion kann e Benachrichtigung an verschiedene Personengruppe gesendet werden.

Die Grundfunktionalität inklusive Erweiterungen der meisten Systeme decken viele generische Abläufe in Forschungsinfrastrukturen gut ab. Sollen spezifische unterstützt und abgebildet werden, sind eigene Modifizierungen und Anpassungen nötig. Aus diesem Grund wird die Verwendung von *Open Source Software* (O

Einbindung neuer Werkzeuge und Dienste

Error creating thumbnail: File missing

Service Life Cycle von der DARIAH-DE Forschungsinfrastruktur, Quelle: https://de.dariah.eu/dariah-svg/small/131202_QS_690.png

Die Integration von neuen Diensten in einer Forschungsinfrastruktur ist ein komplexer Vorgang, bei dem sowohl technische als auch fachwissenschaftliche Anfo berücksichtigt werden müssen. Zu diesem Zweck sollten Dienste bereits während des Entwicklungsprozesses bis zum Produktivbetrieb von Mentoren aus beide betreut und begleitet werden.

Die Abbildung illustriert einen möglichen Ablauf, wie ein neuer fachwissenschaftlicher Dienst in eine Forschungsinfrastruktur (Beispiel DARIAH-DE) aufgeno kann. In der ersten Phase, dem Proposal State, werden Mentoren benannt und es wird evaluiert, ob ein Dienst oder das Werkzeug sich als DARIAH-Komponent Die technischen und fachwissenschaftlichen Mentoren begleiten den kompletten Integrationsprozess. Nach einer Entscheidung beginnen die Entwicklungen in d State. Hier wird der betreffende Dienst und dessen Dokumentation weiterentwickelt und in die DARIAH-Infrastruktur eingebunden. Um eine hohe Qualität sich erfolgen im Anschluss an die Entwicklung ausführliche Tests. In diesem Testing State wird der Dienst von geisteswissenschaftlichen Nutzergruppen getestet und Rückmeldung eingeholt. Bei Bedarf wird der Dienst weiterentwickelt. Damit der entwickelte Dienst in die Produktionsphase übergehen kann, müssen im Hando Komponenten, z.B. Software, Daten und Dokumentation, einem Service Hosting Team übergeben werden. In der letzten Phase, dem Production State, sorgt DA nachhaltigen Betrieb, die Pflege des Dienstes sowie dessen Verbreitung. Der Dienst steht allen Benutzern zur Verfügung, die die DARIAH Terms of Use akzepti

Links und Literatur

Scott Cantor, John Kemp, Rob Philpott, Eve Maler. Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0, http://docs.oas open.org/security/saml/v2.0/saml-core-2.0-os.pdf

Tibor Kálmán, Daniel Kurzawe, Ulrich Schwardmann: European Persistent Identifier Consortium - PIDs für die Wissenschaft. In: Reinhard Altenhöner und Clau (Hrsg.): Langzeitarchivierung von Forschungsdaten – Standards und disziplinspezifische Lösungen, Berlin 2012, S. 151 – 168 ISBN 978-3-944417-00-4

Danah Tonne, Jedrzej Rybicki, Stefan E. Funk, Peter Gietz . Access to the DARIAH Bit Preservation Service for Humanities Research Data. In *21st Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP2013),* S. 9-15

Anmerkungen

1. ↑ https://de.dariah.eu/dariah-visualisiert
2. ↑ https://de.dariah.eu/
3. ↑ http://www.textgridrep.de/
4. ↑ http://etherpad.org/ - die DARIAH Instanz findet man unter https://etherpad.de.dariah.eu/
5. ↑ https://de.dariah.eu/developer-portal
6. ↑ https://www.atlassian.com/software/confluence - die DARIAH Instanz findet man unter https://wiki.de.dariah.eu/
7. ↑ https://listserv.gwdg.de/mailman/listinfo
8. ↑ http://subversion.apache.org/ - die DARIAH Instanz findet man unter http://dev.dariah.eu/svn/repos
9. ↑ https://www.atlassian.com/software/jira - die DARIAH Instanz findet man unter http://dev.dariah.eu/jira
10. ↑ Projektverwaltung, https://projects.gwdg.de/
11. ↑ http://jenkins-ci.org/ - die DARIAH Instanz findet man unter https://ci.de.dariah.eu/jenkins/
12. ↑ Danah Tonne, Jedrzej Rybicki, Stefan E. Funk, Peter Gietz . Access to the DARIAH Bit Preservation Service for Humanities Research Data. In *21st Eur International Conference on Parallel, Distributed and Network-Based Processing (PDP2013), S. 9-15*
13. ↑ http://irods.org/
14. ↑ http://www.pidconsortium.eu/ oder auch: Tibor Kálmán, Daniel Kurzawe, Ulrich Schwardmann: European Persistent Identifier Consortium - PIDs für d In: Reinhard Altenhöner und Claudia Oellers (Hrsg.): Langzeitarchivierung von Forschungsdaten – Standards und disziplinspezifische Lösungen, Berlin 2 168. ISBN 978-3-944417-00-4
15. ↑ http://eudat.eu/ http://eudat.eu/
16. ↑ http://clarin.eu/
17. ↑ https://textgrid.de/
18. ↑ http://www.handle.net/

19. ↑ https://de.dariah.eu/repository
20. ↑ http://www.doi.org/
21. ↑ Scott Cantor, John Kemp, Rob Philpott, Eve Maler. Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0, http://
open.org/security/saml/v2.0/saml-core-2.0-os.pdf
22. ↑ https://shibboleth.net/
23. ↑ https://www.aai.dfn.de/
24. ↑ http://services.geant.net/edugain/Pages/Home.aspx

Die Zukunft im Blick: Nachhaltigkeit und Nachnutzbarkeit

Allzu oft ist es üblich, dass Werkzeuge und Dienste, die im Rahmen von Forschungsprojekten entwickelt werden, nach Ablauf des Förderzeitraums nicht mehr s
weiterentwickelt bzw. gepflegt werden und so anderen WissenschaftlerInnen gar nicht erst zur Nachnutzung angeboten werden können. Eine Möglichkeit ist die
erstellten Software und Erkenntnisse zur Integration in eine Forschungsinfrastruktur. Auf diese Weise haben WissenschaftlerInnen zentrale Anlaufstellen und
Mehrfachentwicklungen können vermieden werden. Der Begriff Nachhaltigkeit lässt sich in verschiedene Teilaspekte unterteilen: Fachwissenschaftliche, daten-
technische und betriebliche- und organisatorische Nachhaltigkeit.

Fachwissenschaftliche Nachhaltigkeit

Wenn es keine ForscherInnen und WissenschaftlerInnen mit dem entsprechenden Verständnis für neue Methoden der Datennutzung, -analyse und -interpretation
selbst umfangreichste Repositorien mit hochwertigen Forschungsdaten letztlich wertlos. Erworbenes Wissen kann jedoch durch vielfältige Angebote aus dem Be
und in einer Forschungsinfrastruktur an die nächste Generation weitergegeben und durch neue Anregungen ergänzt werden. Durch einen Diskurs innerhalb der C
können neue Ansätze kritisch hinterfragt, an spezifische Anforderungen angepasst, weiter verbessert und schließlich fester Teil des Forschungsprozesses werden

Technische Nachhaltigkeit

Auf der einen Seite müssen relevante fachwissenschaftliche Dienste technisch betreut und in die Infrastruktur integriert werden. Zu diesem Zweck kann die betr
Infrastruktur beispielsweise einem Projekt ein Mentorenteam bestehend aus einem technischen und einem fachwissenschaftlichen Experten zur Verfügung stelle
möglicherweise bereits während der Entwicklung Hinweise zur Verwendung bestimmter Standards und Schnittstellen geben. Auf der anderen Seite muss bei der
Basisdiensten ein besonderer Augenmerk auf Modularität und leichter Austauschbarkeit der verwendeten Komponenten gelegt werden. Technologien unterliegen
Wandel und müssen in regelmäßigen Abständen ersetzt werden, was auch in diesem Fall durch die Verwendung von Standards und standardisierten Schnittstelle
wird.

Daten-technische Nachhaltigkeit

Im Kontext der Nachhaltigkeit von Daten ist beispielsweise die technische Interoperabilität von Daten und Werkzeugen sowie der Zugang zu den Daten und der
Langzeitarchivierung zu nennen. Diese beiden Herausforderungen können vor allem durch organisatorische und konzeptuelle Maßnahmen wie generische Stand
Empfehlungen für Prozeduren und Organisationsstrukturen angegangen werden.

Betriebliche und organisatorische Nachhaltigkeit

Der Betrieb einer Forschungsinfrastruktur muss langfristig gesichert werden, auch wenn beispielsweise die beteiligten Organisationen in einigen Jahren nicht m
Auch in der Zukunft müssen Kosten für Ressourcen und Personal abgedeckt und eine dynamische Infrastruktur weiter betreut werden. Nur wenn ForscherInnen
können, dass verwendete Komponenten längerfristig erhalten bleiben, kann Vertrauen in eine Forschungsinfrastruktur entstehen.

Ein besonders auf Nachhaltigkeit ausgelegtes Förderprogramm ist das *European Strategy Forum on Research Infrastructures* (ESFRI)[22], das durch die besond
Projekte eine Förderdauer von mindestens 15 Jahren vorsieht. Im Bereich der Geistes- und Kulturwissenschaften konnten einige Projekte bereits auf der so gena
Roadmap[23] platziert werden.

Anmerkungen

1. ↑ Eine umfangreiche Sammlung von Lehrmaterialien zu allen Bereichen der Digital Humanities findet sich auf https://www.oercommons.org/groups/daria
2. ↑ Unter http://dh-registry.de.dariah.eu/courses/index/country:germany wurde eine Übersicht von Studiengängen der Digital Humanities geschaffen