DH Tag

Philosophische Fakultät II

Anke Lüdeling

Humboldt-Universität zu Berlin

Forschungsdaten an der Philosophischen Fakultät II
1. Treffen | 21.05.2015

- 14:00 - 15:00 Uhr:
  Kurze Einführungen von Anke Lüdeling und Malte Dreyer
- 15:00 - 16:30 Uhr:
  Postersession
- ab 16:30 Uhr:
  Diskussion

1. Forschungsdaten

2. Warum soll man Daten veröffentlichen?
   - Forschung
   - Replizierbarkeit
   - Wiederverwendbarkeit der Daten
   - Anforderung der Geldgeber und Universitäten

3. Technisches zur Veröffentlichung
   - Formate und Architekturen
   - Repositorien

Forschungsdaten

Editionen Quellen (Bilder) und Transliterationen, kritische Apparate etc.

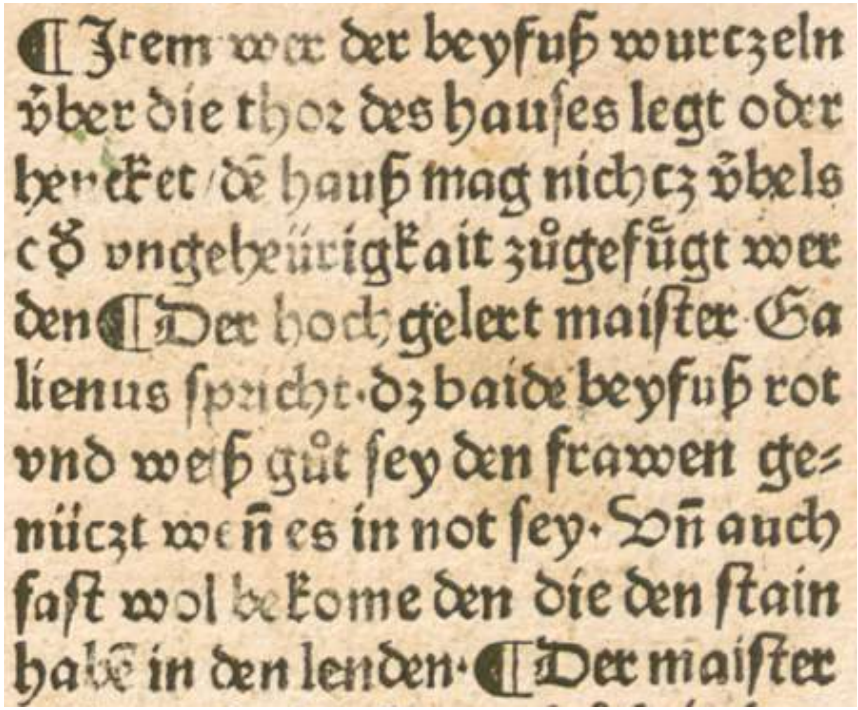
Korpora Textsammlungen (gesprochene oder geschriebene Daten, Videos), mit Metadaten und ggf. Annotationen

Lexika Einträge mit bestimmten Informationen, dazu ggf. Verweise auf Belege

Daten aus Erhebungen und Experimenten Fragebögen, Stimulusmaterial, Messdaten (Reaktionszeiten, Potentiale, etc.)

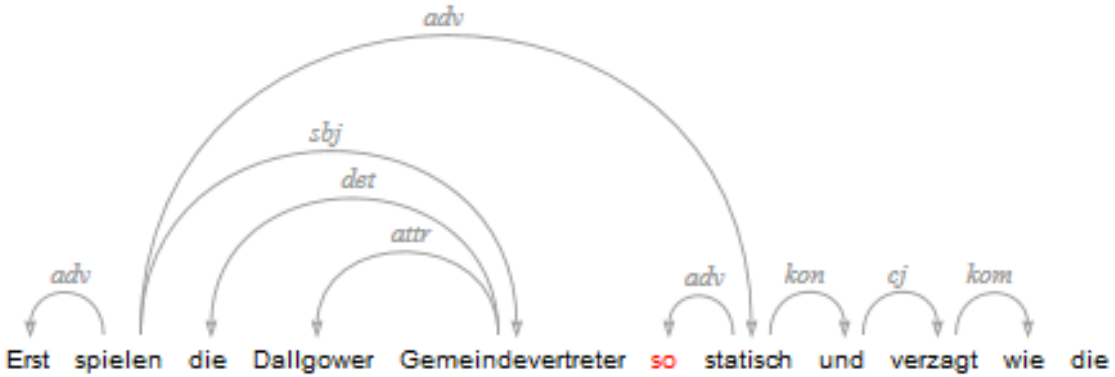
Programme und Skripte Software zur Suche in Daten, zur Verarbeitung & Analyse von Daten

Editionen


¶Jtem wer der beyfuſz wurczeln
v̊ber die thor des hauſes legt oder
hencket / dẽ hauſz mag nichcz v̊bels
od̊ vngeheürigkait zůgefůgt wer
den ¶ Der hochgelert maiſter Ga
lienus ſpricht + dz baide beyfuſz rot
vnd weiſz gůt ſey den frawen ge=
nüczt weñ es in not ſey + Vñ auch
faſt wol bekome den die den ſtain
habẽ in den lenden + ¶ Der maiſter
(Gart der Gesundheit, 1487)

Korpora

Erst	spielen	die	Dallgower	Gemeindevertreter	so	statisch	und	verzagt	wie	die
erst	spielen	der	Dallgower	Gemeindevertreter	so	statisch	und	verzagt	wie	der
--	3.Pl.Pres.Ind	Nom.Pl.Masc	Pos.*.*.*	Nom.Pl.Masc	--	Pos	--	Pos	--	Nom.Sg.Fem
ADV	VVFIN	ART	ADJA	NN	ADV	ADJD	KON	ADJD	KOKOM	ART



Steilpass Wunder gibt es immer wieder ! Erst spielen die Dallgower Gemeindevertreter so statisch und verzagt wie die deutsche Abwehr
von dem man hofft , dass die Seeburger oder Groß-Glienicker Mitspieler ihn aufnehmen können . Ein Befreiungsschlag ist es allerdings

Experimentelle Daten

	Trial	h	beding	code	RT	group
1	1	27	130	170	508	EAHP
6554	254	167	130	170	311	EAHP
13108	288	78	131	180	383	Kontrolle
19662	192	223	140	170	412	LAHP
26216	96	270	141	181	806	LALP
32770	350	288	143	180	343	LALP

Tabelle: Reaktionszeitdaten (Dank an Felix Golcher und Juliane Domke)

Forschungsdaten in den Geisteswissenschaften

Forschungsdaten in den Geistewissenschaften können sehr unterschiedlich sein.

Aufbereitungszustand: Primärtext, Annotationen, Verweise auf externe Ressourcen | abgeschlossen oder erweiterbar

Zweck: Forschung, Lehre | Datengrundlage, Verarbeitung

Quelle: historisch, modern | handschriftlich, gedruckt, digital | Bild, Text, Audiodatei, Videodatei

Größe: ein Text, Internetkorpora

...

1. Forschungsdaten

2. Warum soll man Daten veröffentlichen?
   - Forschung
   - Replizierbarkeit
   - Wiederverwendbarkeit der Daten
   - Anforderung der Geldgeber und Universitäten

3. Technisches zur Veröffentlichung
   - Formate und Architekturen
   - Repositorien

Warum soll man Forschungsdaten veröffentlichen?

Oder auch: Reicht es nicht, die Forschungsergebnisse zu veröffentlichen?

- Erstellung von Ressourcen ist Forschung
- Nachvollziehbarkeit der Ergebnisse (Replizierbarkeit)
- Wiederverwendbarkeit der Daten
- Anforderungen der Geldgeber

1. Forschungsdaten

2. Warum soll man Daten veröffentlichen?
   - Forschung
   - Replizierbarkeit
   - Wiederverwendbarkeit der Daten
   - Anforderung der Geldgeber und Universitäten

3. Technisches zur Veröffentlichung
   - Formate und Architekturen
   - Repositorien

Erstellung von Ressourcen

In die Erstellung von qualitativ hochwertigen Ressourcen fließt viel Forschung und gute Ressourcen ermöglichen weitere Forschung.

Gut entworfene und nutzbare Ressourcen sind veröffentlichungswürdig - es gibt Konferenzen & Zeitschriften für die Veröffentlichung von Ressourcen. Ressourcenveröffentlichungen werden viel zitiert!

Transparenz und Replizierbarkeit

Bei der Erstellung *jeder* digitalen Ressource wird interpretiert.

Erstellung der Primärebene: Entscheidungen über Spatien, Entscheidungen über Wiedergabe von Zeichen, Entscheidungen über Wiedergabe von Audiodaten etc.

Erstellung von Annotationen (Kategorisierungen): Entscheidungen über Kriterien, Entscheidungen über Werkzeuge, Evaluationen etc.

Erstellung von Verweisen: Entscheidungen über die externen Quellen

Analyse: Entscheidungen über Analyseverfahren, statistische Verfahren etc.

...

Das bedeutet, dass Ergebnisse nur nachvollziehbar (und ggf. replizierbar) sind, wenn alle Daten veröffentlicht werden.

Wiederverwendbarkeit und Nachnutzung

Die Erstellung von digitalen Ressourcen ist oft teuer und aufwändig. Die Daten können in vielen Fällen für weitere Forschungsfragen genutzt werden.

- innerhalb eines Faches: z.B. können mehrere (unabhängige) linguistische Forschungsfragen auf derselben Textgrundlage behandelt werden
- fächerübergreifend: z.B. kann ein historischer Text, der in der Geschichtswissenschaft digitalisiert wird, für literaturwissenschaftliche oder linguistische Zwecke ausgewertet werden

Datenaufbewahrung

"Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden."

Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“;

Denkschrift DFG, 1998

Datenaufbewahrung

"Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden."

Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“;

Denkschrift DFG, 1998

"Wenn aus Projektmitteln systematisch Forschungsdaten oder Informationen gewonnen werden, die für die Nachnutzung durch andere Wissenschaftlerinnen und Wissenschaftler geeignet sind, legen Sie bitte dar, ob und auf welche Weise diese für andere zur Verfügung gestellt werden. Bitte berücksichtigen Sie dabei auch - sofern vorhanden - die in Ihrer Fachdisziplin existierenden Standards und die Angebote existierender Datenrepositorien oder Archive."

DFG-Vordruck 54.01; Leitfaden für die Antragstellung Projektanträge [06/14]

1. Forschungsdaten

2. Warum soll man Daten veröffentlichen?
   - Forschung
   - Replizierbarkeit
   - Wiederverwendbarkeit der Daten
   - Anforderung der Geldgeber und Universitäten

3. Technisches zur Veröffentlichung
   - Formate und Architekturen
   - Repositorien

Technisches zur Veröffentlichung

- Abklärung der **Rechte** (Copyright und Persönlichkeitsrechte), z. B. durch CLARINs Legal Help Desk `http://de.clarin.eu/en/training-helpdesk/legal-helpdesk.html`
- genaue **Dokumentation** der Daten erforderlich (dies betrifft die Erstellung und alle Verarbeitungsschritte, insbesondere auch Evaluationen)
- **Versionierung** des Korpus, Bezugnahme auf die verwendete Version, Aufbewahrung alter Versionen
- freie Veröffentlichung des Korpus unter einer passenden **Lizenz**:
  Daten und Programme sollten immer unter einer möglichst offenen Lizenz abgelegt werden, sonst können sie nicht (gefahrlos) weiter genutzt werden.
  Für Daten sind oft die Creative Commons-Lizenzen nützlich, für Programme beispielsweise die Gnu Public License oder Apache-Lizenzen (siehe `http://opensource.org/licenses`).

Open Source Daten und freie Software

Frei wie „für umsonst“. Nicht jedeR hat ausreichend Geld oder den Zugang zu Campuslizenzen.

Frei wie „offen und transparent“. EinE WissenschaftlerIn sollte die Möglichkeit haben, nachzuvollziehen, wie eine Ressource aufgebaut wurde oder was ein Programm tut.

Frei wie „nicht fest“: Freie Software und freie Ressourcen sind durch andere ForscherInnen modular erweiterbar.

- für die Statistiksoftware `R` derzeit 6387 Erweiterungen (1. März 2015 17:05:46)
- unser Lernerkorpus Falko wird von vielen Studierenden für Abschlussarbeiten genutzt und dabei ständig erweitert

Formate und Architekturen

Formate Die verwendeten Formate sollten gut beschrieben und weit verbreitet sein. Es ist sinnvoll, die Daten in mehreren Formaten abzulegen - eines davon möglichst ein XML-Format (für gut beschriebene Formate gibt es oft Konvertierer, zum Beispiel SaltNPepper `http://korpling.german.hu-berlin.de/saltnpepper/`). Für viele Ressourcentypen gibt es etablierte (zertifizierte oder de facto) Standards (beispielweise von der Text Encoding Initiative `http://www.tei-c.org/index.xml` oder von der ISO).

Architekturen Die Daten sollten in Architekturen abgelegt werden, die offen und erweiterbar sind. Auch hier ist es sinnvoll, bereits existierende Lösungen zu nutzen und ggf. zu erweitern.

1. Forschungsdaten

2. Warum soll man Daten veröffentlichen?
   - Forschung
   - Replizierbarkeit
   - Wiederverwendbarkeit der Daten
   - Anforderung der Geldgeber und Universitäten

3. Technisches zur Veröffentlichung
   - Formate und Architekturen
   - Repositorien

Repositorien

In Repositorien werden Daten aufbewahrt und - durch die Angabe von **Metadaten** - suchbar gemacht. Es gibt ein Kontinuum von sehr allgemeinen Repositorien, in denen alle möglichen Daten abgelegt werden (wie bspw. das Virtual Language Observatory `https://www.clarin.eu/content/virtual-language-observatory`), bis hin zu sehr spezifischen Repositorien, in denen nur Daten eines bestimmten Typs abgelegt werden (wie bspw. das LAUDATIO-Repositorium `http://www.laudatio-repository.org/repository/` für historische Texte).

Problem: Es gibt bisher für geisteswissenschaftliche Forschungsdaten keine **langfristig gesicherten** Repositorien, die mehr versprechen als die reine 'Aufbewahrung':

- Daten können nicht verändert werden
- Programme, mit denen die Ressourcen gelesen und analysiert werden können, werden nicht notwendigerweise gewartet

Zusammenfassung

Forschungsdaten An der Fakultät gibt es viele digital vorliegende Forschungsdaten in ganz unterschiedlichen Formaten.

Veröffentlichung Es gibt gute Gründe, digitale Ressourcen frei zu veröffentlichen. Für die technischen Aspekte (Formate, Rechte) kann man Lösungen finden.

Langfristige Speicherung & Zugänglichkeit Wir brauchen die Unterstützung der Universität bei der nachhaltigen Speicherung und der Sicherung der Zugänglichkeit.

Forschungsdatenmanagement

„Die in einem System integrierte Verarbeitung und Darstellung heterogener Datenmengen aus ganz unterschiedlichen Quellen, erfasst mit verschiedenen Instrumenten, erfordert einen langen Prozess an Transformation, Speicherung und Übermittlung. Dieser Prozess muss bewusst und nachvollziehbar gestaltet werden, damit die erzeugten Daten ihre wissenschaftliche Aussagekraft behalten und für die Auswertung zugänglich bleiben. Das ist die Aufgabe des Forschungsdatenmanagements. Es findet in dem Bewusstsein statt, dass lokale Lösungen Bestandteil einer übergreifenden Forschungsdateninfrastruktur sein müssen. Das Forschungsdatenmanagement muss so gestaltet werden, dass Datenzugriff und -auswertung unabhängig vom Datenerzeuger möglich wird und bleibt.“

Büttner, St., Hobohm, H. & Müller, L. (Hg.) Handbuch Forschungsdatenmanagement. Bock und Herchen Verlag. Bad Honnef. S.29

Ziele

Vernetzung innerhalb der Fakultät und gerne auch mit KollegInnen aus anderen Fakultäten, die mit ähnlichen Daten arbeiten (Geschichte, Musikwissenschaft, Psychologie etc.).

Forschungsdatenstrategien der Fakultät entwickeln und Forschungsdatenstrategien der Universität beeinflussen (Forschungsdaten gibt es nicht nur in den Naturwissenschaften)

Eine Liste aller angemeldeten Poster finden Sie unter `http://linguistik.hu-berlin.de/dh-tag`.
Wenn Sie möchten, können Sie uns das pdf für Ihr Poster schicken und gerne auch einen Link zu Ihrem Projekt.