Datenmanagementplan

zur Masterarbeit „Modelling and Automated Retrieval of Provenance Relationships“

Thomas Schneider

12. Juni 2023

1 Allgemein

1.1 Thema

1.1.1 Wie lautet die primäre Forschungsfrage der Abschlussarbeit?

Wie können Provenienzbeziehungen modelliert und maschinell gestützt aufgefunden werden?

1.1.2 Bitte geben Sie einige Schlagwörter zur Forschungsfrage bzw. Fragestellung an.

- DDC:[1]
  - 005.72 Datenaufbereitung und Datenrepräsentation
  - 006.332 Wissensrepräsentation
  - 020.0113 Computermodellierung in Bibliotheks- & Informationswissenschaften
- 2012 ACM Computing Classification System:[2]
  - Information systems / Information retrieval / Retrieval tasks and goals / Question answering
  - Information systems / Information retrieval / Retrieval tasks and goals / Information extraction

1.1.3 Welchen Regeln oder Richtlinien (HU) zum Umgang mit den in der Abschlussarbeit erhobenen Forschungsdaten folgen Sie für den DMP? Bitte referenzieren Sie diese hier inklusive Version bzw. Veröffentlichungsjahr.

Institut für Bibliotheks- und Informationswissenschaft: Leitlinie zum Umgang mit Forschungsdaten in Abschlussarbeiten. Beschlossen im Institutsrat des IBI am 08.12.2021, in Kraft getreten am 01.02.2022.[3]

[1]↑ https://deweysearchde.pansoft.de/webdeweysearch/mainClasses.html?catalogs=DNB
[2]↑ https://dl.acm.org/ccs
[3]↑ https://www.ibi.hu-berlin.de/de/studium/rundumdasstudium/fdm-fuer-studierende

2 Inhaltliche Einordnung

NB: Bitte beschreiben Sie jeden Datensatztyp oder Datensammlung einzeln in dem jeweiligen Kapitel, wo sinnvoll.

2.1 Datensatz

2.1.1 Um welche Arten von Daten handelt es sich? Bitte in wenigen Zeilen kurz beschreiben.

Für die Literaturstudie und Analyse der Datenquellen wurden folgende Daten gesammelt:

(1) bibliographische Metadaten zu relevanten Arbeiten aus der Literatur

(2) für ausgewählte Arbeiten: Volltexte im PDF-Format

(3) Metadaten von relevanten Webseiten

(4) statistische Angaben zu Datenquellen, z. B. die Anzahl der darin enthaltenen Datensätze

2.2 Datenursprung

2.2.1 Werden die Daten selbst erzeugt oder nachgenutzt?

(1) teilweise nachgenutzt aus Portalen, teilweise selbst erzeugt

(2) heruntergeladen von Verlags- und Aggregatorplattformen

(3) selbst erzeugt

(4) aus der Literatur und von Webseiten selbst extrahiert

2.2.2 Wenn die Daten nachgenutzt werden, wer hat die Daten erzeugt? Bitte mit Angabe des Identifiers, falls vorhanden, z.B. DOI.

Volltexte und deren Metadaten wurden von diversen Portalen heruntergeladen, darunter:

- https://hu-berlin.hosted.exlibrisgroup.com/primo-explore/search?vid=hub_ub
- https://www.sciencedirect.com/
- https://ebookcentral.proquest.com/lib/huberlin-ebooks/home.action
- https://www.tandfonline.com/
- https://dblp.uni-trier.de/
- https://gvk.k10plus.de

2.3 Reproduzierbarkeit

2.3.1 Sind die Daten reproduzierbar, d. h. ließen sie sich, wenn sie verloren gingen, erneut erstellen oder erheben?

Die Daten sind nur bedingt reproduzierbar: Anhand des Literaturverzeichnisses der Arbeit lassen sich alle Quellen wiederfinden, aber bei den meisten Webseiten besteht die Gefahr, dass

der Inhalt sich ändert oder die Webseite inaktiv wird. Letzteres kann auch bei denjenigen der wissenschaftlichen Arbeiten passieren, die nur online verfügbar sind.

2.4 Nachnutzung

2.4.1 Für welche Personen, Gruppen oder Institutionen könnte dieser Datensatz (für die Nachnutzung) von Interesse sein? Für welche Szenarien ist dies denkbar?

Der Datensatz könnte von Interesse sein für Forschende, die sich mit derselben Thematik befassen und einige Aspekte aus oder neben dieser Arbeit vertiefen möchten. Er könnte auch als (unvollständige) Basis für einen Übersichtsartikel über bibliographische Datenquellen oder digitale Provenienzforschung dienen.

3 Technische Einordnung

3.1 Datenerhebung

3.1.1 Wann erfolgt(e) die Erhebung bzw. Erstellung der Daten?

Die Sammlung der Daten erfolgte über gesamten Bearbeitungszeitraum der Masterarbeit hinweg, d. h. vom 16.02.2023 bis 14.06.2023.

3.1.2 Wann erfolgt(e) die Datenbereinigung/-aufbereitung bzw. Datenanalyse?

(1) Bei jedem Fund einer wissenschaftlichen Arbeit wurden die bibliographischen Metadaten von der entsprechenden Plattform heruntergeladen und eine Masterdatei importiert (bzw. von Hand eingetragen, falls nötig). Die Daten wurden bei jedem Import sofort geprüft und ggf. korrigiert. Eine weitere Datenbereinigung oder -aufbereitung ist nicht erforderlich.

(2) Die elektronisch vorliegenden wissenschaftlichen Arbeiten, von denen abzusehen war, dass sie im Bearbeitungszeitraum der Masterarbeit länger oder mehrfach konsultiert werden mussten, wurden als PDF-Datei gespeichert. Eine Datenbereinigung oder -aufbereitung scheint hier nicht sinnvoll.

(3) Wie (1); hier wurden alle Metadaten von Hand eingegeben.

(4) Die statistischen Daten der Datenquellen wurden am 09.06.2023 in einer Datei gebündelt.

3.2 Datengröße

3.2.1 Was ist die tatsächliche oder erwartete Größe der Daten(typen)?

Stand 09.06.2023:

(1) eine Datei zu 115 kB

(2) ca. 150 Dateien zu insgesamt 412,8 MB

(3) in (1) enthalten

(4) eine Datei zu 1,4 kB

3.3 Formate

3.3.1 In welchen Formaten liegen die Daten vor?

(1) .bib (BibTeX)

(2) .pdf (Portable Document Format)

(3) siehe (1)

(4) .md (Markdown)

3.4 Werkzeuge

3.4.1 Welche Instrumente, Software, Technologien oder Verfahren werden zur Erzeugung, Erfassung, Bereinigung, Analyse und/oder Visualisierung der Daten genutzt? Bitte (falls möglich) mit Versionsnummer und Referenz in Form einer Adresse jeweils angeben.

(1) *Download:*
Firefox 113.0.2 (64-bit) https://www.mozilla.org/en-GB/firefox/new/

Verwaltung:
JabRef 5.9–2023-01-08–76253f1a7 https://www.jabref.org/

(2) *Download:* wie (1)

Betrachtung:
Skim 1.6.16 (146) https://skim-app.sourceforge.io/
Adobe Acrobat Reader 2023.001.20177 https://www.adobe.com/

(3) siehe (1)

(4) MacDown 0.7.3 (1008.4) https://macdown.uranusjr.com/
Obsidian 1.3.5 (Installer 0.15.9) https://obsidian.md/

3.4.2 Welche Software, Verfahren oder Technologien sind notwendig, um die Daten zu nutzen?

(1) mindestens: Texteditor; besser: Literaturverwaltungsprogramm

(2) PDF-Reader wie z. B. Adobe Acrobat Reader oder Preview (MacOS)

(3) siehe (1)

(4) Texteditor oder Markdown-Editor

3.5 Versionierung

3.5.1 Werden verschiedene Versionen der Daten erzeugt (z. B. durch verschiedene Weiterbearbeitungsprozesse bzw. Bereinigung von Daten)?

*** TBC ***