Exploration und Visualisierung bibliographischer Metadaten der BDSL im GiNLab - Ein Projektbericht

von Michel Schwab

11.08.2021

1 Einleitung

Bibliographische Metadaten sind oft frei verfügbar, allerdings stehen meist keine Werkzeuge zur Exploration oder Visualisierung der Daten zur Verfügung. Im Folgenden wird ein Prototyp eines solchen Werkzeugs vorgestellt, der im Rahmen des GiNLabs entstanden ist. Ausgangspunkt für die Einrichtung des GiNLabs war 2017 die DFG-Beantragung eines Fachinformationsdienstes (FID) Germanistik mit dem Ziel, das Portal „Germanistik im Netz (GiN)“ zum zentralen Einstiegspunkt für das Fach Germanistik auszubauen. Als Leitplanken dienen dabei vier wesentliche Aktivitäten philologischer Praxis: Recherchieren, Publizieren, Informieren, Forschen. Im Modul „Forschen“ (https://www.germanistik-im-netz.de/forschen/) wurde das GiNLab eingerichtet, um dort ein Data-Mining-Werkzeug zur Analyse bibliographischer Metadaten in enger Abstimmung mit der Zielgruppe zu entwickeln. Zur Verfügung standen dafür rund 490.000 Titeldatensätze aus dem Zeitraum 1985-2019 aus der wichtigsten germanistischen Fachbibliographie, der Bibliographie der deutschen Sprach- und Literaturwissenschaft (BDSL).
Wissenschaftler*innen sollen durch das Werkzeug bei der Nutzung und Auswertung von Informationen und Daten, insbesondere bibliographischer Metadaten, unterstützt werden. Dabei sollen sie Aufschlüsse darüber gewinnen können, wie sich germanistische Arbeitsfelder, Methoden und personelle Netzwerke in den vergangenen Jahrzehnten gewandelt haben. Der Schwerpunkt liegt dabei auf Datenvisualisierungen, welche in einem webbasierten, interaktiven Prototyp umgesetzt wurden.
Als Projektpartner der Goethe-Universität Frankfurt/M. konnte das Institut für Bibliotheks- und Informationswissenschaft der HU Berlin mit der Professur für „Information Processing and Analytics“ (Herr Prof. Dr. Robert Jäschke und Michel Schwab) gewonnen werden. Die Projektlaufzeit wurde auf 1,5 Jahre festgelegt, beginnend am 01.01.2020. Das Projekt wurde in drei Phasen aufgeteilt: Ideenfindung, Implementierung der Visualisierungen und Erstellung eines Prototyps. Besonderes Augenmerk lag dabei auf der methodischen Umsetzung der Phase Ideenfindung, die im Rahmen eines Projektseminars gemeinsam mit dem Institut für Bibliotheks- und Informationswissenschaft (IBI) der Humboldt-Universität zu Berlin umgesetzt wurde. Der im Anschluss implementierte Prototyp wurde im Rahmen eines Workshops gemeinsam mit der Fachcommunity weiterentwickelt.
In diesem Beitrag wird zunächst die methodische Umsetzung der Ideenfindung vorgestellt und anschließend die technische Umsetzung des Prototyps beschrieben.

2 Ein Projektseminar zur Ideenfindung

Das Seminar diente dem Ziel, Mockups von interaktiven Datenvisualisierungen zu entwickeln, die die Beantwortung (proto)typischer wissenschaftlicher Fragestellungen unterstützen können.
Dazu wurde zunächst ein Konzept entwickelt, wie die Bachelor- und Masterstudierenden des Instituts für Bibliotheks- und Informationswissenschaft in das Projekt einbezogen werden konnten. Im Sommersemester 2020 wurde ein Projektseminar von Michel Schwab angeboten, in dem Studierende in Kleingruppen von circa drei Personen wissenschaftliche Fragestellungen erarbeiten und diese mit Hilfe von Visualisierungen der BDSL-Metadaten beantworten sollten.
Insgesamt gab es drei Phasen:

1. Ideenfindung (inklusive Feedback-Runde)
2. Konzeptausarbeitung (inklusive Feedback-Runde)
3. Implementierung und Hausarbeit

Während der Ideenfindung haben die Kleingruppen die Daten exploriert, zusammen erste Fragestellungen ausgearbeitet und erste Skizzen angefertigt, um diese mit Hilfe einer Visualisierung zu beantworten. Am Ende der Phase gab es eine Feedback-Runde, damit die Studierenden Unterstützung und neue Ideen zur Verfeinerung ihrer Ideen bekommen. Dazu wurden zwei Wissenschaftler*innen aus dem Institut für deutsche Literatur an der HU Berlin eingeladen, die sich mit den Studierenden per Videokonferenz getroffen und ihre Ideen diskutiert haben. Dadurch gab es einen ersten Austausch mit der Fachcommunity. Nach dieser Runde gab es zusätzlich eine Peer-Feedback-Runde zwischen den einzelnen Gruppen, um Verbesserungen und weitere Ideen auszutauschen.
In der zweiten Phase des Seminars wurde sich innerhalb der Gruppen auf eine oder zwei Ideen geeinigt und ein Konzept angefertigt. Hier galt es zu entscheiden, auf welche Ideen man sich konzentriert, welche Daten man benötigt und wie man die Visualisierung technisch umsetzen kann. Am Ende dieser Phase gab es eine Zwischenpräsentation der Ergebnisse inklusive einer zweiten Feedback-Runde mit den Wissenschaftler*innen.
Nun kam die letzte Phase des Seminars, in der die Studierenden ihr ausgearbeitetes Konzept durch das Feedback verfeinern und umsetzen mussten. Ein Fokus lag auf der technischen Umsetzbarkeit, ein zweiter auf der Benutzerfreundlichkeit. Welche Tools stehen zur Verfügung? Ist es möglich, alle benötigten Informationen und Daten zu erhalten? Welche Analysen müssen gemacht werden, bevor die Visualisierung umgesetzt werden kann? Wie soll der Nutzer mit dem Mockup interagieren? Die Studierenden haben verschiedene Visualisierungs- oder Mockup-Tools getestet, um ihre Idee zu skizzieren und umzusetzen. Zudem wurde ein Portfolio angefertigt, welches am Ende des Seminars präsentiert wurde.
Das Ziel war nicht, einen interaktiven, webbasierten Prototyp im Seminar zu erschaffen, sondern verschiedene Ideen und deren detaillierte Ausarbeitung zu erarbeiten. Deswegen wurde von den Studierenden auch nicht verlangt, ihr Mockup in JavaScript zu implementieren, sondern es durfte frei gewählt werden.

Ergebnisse
Insgesamt gab es sechs Gruppen, die an unterschiedlichen Ideen und Visualisierungen gearbeitet haben. Die Ergebnisse wurden anschließend innerhalb des Projekts diskutiert und zusammen wurde entschieden, welche der Ideen im Prototyp umgesetzt werden sollten.

3 Beschreibung des Prototyps

Datenlage
Die Datenbank BDSL Online umfasste zum Zeitpunkt des Projektstarts rund 490.00 Titeldatensätze. Die Datensätze werden vornehmlich per Autopsie erstellt. Die meisten Einträge sind Monographien oder Aufsätze und Artikel aus Sammelwerken und Zeitschriften. Die BDSL ist in 19 Hauptkategorien eingeteilt. Jeder Titeldatensatz enthält eine große Anzahl an bibliographischen Metadaten, von denen neben einigen typischen Einträgen (Datum der Ersterfassung, Titel, Autor*in, Ko-Autor*in) auch BDSL-spezifische Einträge angegeben sind (Schlagwörter, Haupt- und Unterkategorien der BDSL oder Rezensionsabhängigkeiten).

(Vor)verarbeitung der Daten
Die Daten standen in einem Standard-Datenformat zur bibliothekarischen Katalogisierung (PICA XML) zur Verfügung, welches zuerst in ein für die Verarbeitungssoftware lesbares Format umgewandelt werden musste. Dazu wurde XSLT genutzt, eine Programmiersprache, die XML-Daten transformieren kann.
Die transformierten Daten konnten nun mit dem Catmandu Toolkit weiterverarbeitet werden. Durch das Toolkit konnten die für das Projekt relevanten Metadaten effizient extrahiert und in ein JSON-Format umwandelt werden.
Anschließend wurden verschiedene selbstgeschriebene Python-Skripte zur Datenanalyse genutzt, welche die Daten für die jeweiligen Problemstellungen und Visualisierungen passend verarbeiteten und aggregierten.

Technische Umsetzung des Prototyps
Die einzelnen interaktiven Visualisierungen wurden mit der JavaScript-Bibliothek d3.js umgesetzt. Die Webseite wurde mit Jekyll gebaut, einem Generator für statische Webseiten. Da es sich um einen Prototyp handelt, sind die Visualisierungen noch nicht direkt mit der BDSL Online-Datenbank verbunden, sondern sie enthalten statisch extrahierte Daten.
Die auf der Webseite gezeigten Visualisierungen werden im Folgenden kurz erläutert.

Ergebnisse
1. Geschlechterverteilung in der BDSL (Autor*innen)

Eine interessante Fragestellung, zu der die Studierenden im Projektseminar gekommen sind, war die Frage, wie hoch der Anteil an Autorinnen in der BDSL ist.
Da keine Angaben in den Metadaten existieren, wurden verschiedene Tools ausprobiert, die die Vornamen der Autor*innen in zwei Kategorien (männlich, weiblich) einteilen. Um die Nachnutzbarkeit sicherzustellen, wurde sich für das Open Source Tool "gender" (Python: gender-guesser) entschieden. Auf Basis dieser automatisierten Geschlechterzuordnung der Vornamen der Autor*innen wurden in verschiedenen Szenarien der Anteil an Autorinnen in der BDSL analysiert und durch Linien- und Balkendiagramme wiedergegeben.
Dem Verfasser des Projektberichts ist bewusst, dass die binäre Einteilung von Vornamen problematisch ist; er sieht es aber als einen gelungenen Anfang für ähnliche und tiefergehende Analysen an.

2. Analyse der BDSL-Kategorien

Jeder Titeldatensatz wird bei der Erfassung einer oder mehrerer Kategorien innerhalb der BDSL zugeordnet. Eine zentrale Fragestellung war die zeitliche Entwicklung der 19 Hauptkategorien.

Dies wurde durch einen sogenannten “Streamgraph” dargestellt. Zudem werden die am häufigsten angegebenen Primärautor*innen je Kategorie und Jahr angezeigt. Das Grobgerüst der 19 BDSL-Kategorien, oder besser gesagt, die BDSL-Systematik stammt noch von den beiden Gründungsherausgebern Hanns W. Eppelsheimer und Clemens Köttelwesch. Die Klassifikationsstellen sind aber keineswegs unveränderlich, sondern wurden in den vergangenen sechs Jahrzehnten immer wieder modifiziert und erweitert, zuletzt 2004. Für nicht wenige kanonische und damit häufig behandelte Werke aus dem Mittelalter lassen sich keine Autor*innen im heutigen Sinne angeben; für diese wurde daher der Werktitel gewählt (z.B. Nibelungenlied, Hildebrandslied). Dass gelegentlich die Kurzformen mittelalterlicher Autor*innen auftauchen (z.B. Walther, Wolfram) irritiert; diese Uneinheitlichkeit ist leider nicht durchweg konsequenten Erfassungsrichtlinien geschuldet.

3. Publikationsbeziehungen in Sammelbänden

Eine weitere Idee war die Visualisierung von Publikationsbeziehungen. Da es nur wenige richtige Ko-Autor*innenschaften in der BDSL gibt, wurde eine Idee aus [1] aufgegriffen: Ko-Publikationsbeziehungen zwischen Autor*innen in Sammelbänden. Die Idee basiert auf der Annahme, dass zwei Autor*innen sich kennen, falls beide in demselben Sammelband einen Aufsatz publiziert haben. Aufgrund dieser Annahme wurde ein Netzwerkgraph erstellt, wobei jede*r Autor*in durch einen Knoten dargestellt wird und zwei Autor*innen durch eine Kante verbunden werden, falls die vorher definierte Annahme zutrifft.

In einem weiteren Schritt können nun Communitys in dem Netzwerkgraphen durch den Girvan–Newman-Algorithmus berechnet werden. In einer Community werden Autor*innen zusammengefasst, die untereinander besonders viele Verbindungen aufweisen, d.h. besonders häufig in denselben Sammelbänden Aufsätze publiziert haben. Dies zeigt die Visualisierung. Die Beschriftung der Communitys basiert auf der am häufigsten erfassten Hauptkategorie der Publikationen der Autor*innen einer Community. Die Größe des Knotens ist abhängig von der Anzahl an Autor*innen in der Community.

Die Kanten zwischen den Communitys spiegeln die gemeinsame Publikation von Autor*innen beider Communitys in Sammelbänden wider. Die Breite einer Kante spiegelt die Anzahl der Ko-Publikationen wider. Zusätzlich kann man einen Schwellwert einstellen, der die Mindestanzahl an Ko-Publikationen angibt, damit eine Kante sichtbar ist. Je höher der Wert, desto weniger Kanten werden angezeigt und desto besser kann man die Hauptverbindungen analysieren. In einem nächsten Schritt könnte man nun die einzelnen Communitys analysieren und zum Beispiel aufzeigen, welche Autor*innen besonders viele Verbindungen aufweisen und somit (wahrscheinlich) gut vernetzt sind.

4. Geografische Verortung von Hochschulschriften

Die Idee war es, Wissenschaftler*innen einen geografischen Überblick aller Publikationen in der BDSL zu geben. Dies ist aber unmöglich, da meist nur die Verlagsorte in den Metadaten angegeben sind, die keinen Mehrwert für eine solche Analyse aufweisen. Deswegen wurde sich auf Hochschulschriften (im Wesentlichen Dissertationen und Habilitationen) konzentriert, da dort die benötigten Orts- oder Hochschulangaben in den Metadaten zur Verfügung stehen. Im Hintergrund wird eine Deutschlandkarte angezeigt. Durch Kreise in verschiedenen Größen werden die Anzahl der Publikationen je Stadt dargestellt. Je größer der Kreis, desto mehr

Hochschulschriften wurden hier veröffentlicht. Durch Anklicken der Kreise kann man zusätzlich Informationen sowie die Links zu den Titeldatensätzen in der BDSL Online erhalten.

5. Analyse von Schlagwörtern in der BDSL

In der BDSL werden Titeldatensätze bei der Erfassung verschlagwortet. Eine interessante Fragestellung ist es, herauszufinden, wie sich Schlagwörter zeitlich entwickeln. Welche Schlagwörter wurden seit Anfang an genutzt, welche wurden erst ab einem bestimmten Zeitpunkt vergeben? Es wurde sich dafür entschieden, ein Liniendiagramm zu erstellen, um diese zeitliche Entwicklung visuell darzustellen. Da jede Linie den Verlauf eines Schlagworts darstellt, musste sich auf eine manuelle Auswahl von Schlagwörtern konzentriert werden, da das Diagramm sonst nicht mehr lesbar gewesen wäre. Durch Anklicken in der Legende kann man verschiedene Schlagwörter hervorheben oder aus dem Diagramm entfernen bzw. hinzufügen, um einen besseren Überblick zu erhalten.

4 Ausblick

Der Prototyp ist zurzeit statisch und nicht an eine Datenbank angebunden. Eine Verbesserung wäre die direkte Anknüpfung an die BDSL Online-Datenbank, sodass die Visualisierungen automatisiert mit jedem neu hinzugefügten BDSL-Titeldatensatz erweitert werden würden.
Eine weitere Ausweitung des Prototyps ist die detaillierte Analyse mancher Visualisierungen, zum Beispiel den Community-Graphen. Hier könnte analysiert werden, welche Autor*innen in den einzelnen Communitys vertreten sind und ob es besonders stark vernetzte Autor*innen gibt, die besonders viele Autor*innen miteinander verbinden.
Zudem wurden noch nicht alle Metadaten der BDSL ausgenutzt. Eine Möglichkeit wäre zum Beispiel, Rezensionen in Form eines gerichteten Netzwerkgraphen zu analysieren.
Als einen weiteren Schritt könnte man durch Umfragen analysieren, wie groß das Interesse der Wissenschaftler*innen an solchen Statistiken und Visualisierungen ist, um herauszufinden, in welche Richtung sich ein solcher Prototyp entwickeln soll, und um große Teile der Fachcommunity anzusprechen.

5 Fazit

Durch den Prototyp kann man sehen, dass sich interessante Fragestellungen durch die Visualisierung bibliographischer Metadaten beantworten und analysieren lassen. Der Prototyp ist noch nicht vollständig ausgereift, viele Ideen lassen sich weiter verfeinern und ausweiten. Aber dies ist ein vielversprechender Schritt, um Fachcommunitys einen weiteren Zugang zur Analyse und zum Überblick von bibliographischen Metadaten zu geben.

Die Einbeziehung von Studierenden in das Forschungsprojekt ist eine erfolgreiche Methode zur Ideenfindung von Datenanalysen. Auf der einen Seite lernen Studierende reale Szenarien und Projekte kennen und können sich praxisnah damit auseinandersetzen. Sie lernen zusätzlich neue Tools und Anwendungen kennen. Auf der anderen Seite hat es das Projekt vorangebracht, da viele der Ideen aus dem Seminar als Prototyp umgesetzt werden konnten.

Bibliographie

[1] Kreutel, Jörn; Martus, Steffen; Thomalla, Erika und Zimmer, Daniel. "Die Germanistik der Germanistik" *Internationales Archiv für Sozialgeschichte der deutschen Literatur*, vol. 44, no. 2, 2019, pp. 302-379. https://doi.org/10.1515/iasl-2019-0015