Visualisierungsanwendungen für Bibliotheken und Wissenschaft am Beispiel von *DNBVIS_frodiss*

1. Einleitung

Der folgende Praxisbericht beschreibt die Entwicklung der Visualisierungsanwendung *DNBVIS_frodiss*[1], die das Datenset „Freie Online Hochschulschriften" der Deutschen Nationalbibliothek (DNB) visualisiert[2]. Der Fokus des Praxisberichts liegt auf dem methodischen Vorgehen. Beschrieben werden die Extraktion der benötigten Informationen aus den im MARC21-xml-Metadatenformat vorliegenden Datensätzen sowie die weiteren Verarbeitungsschritte, um die Daten für verschiedene Visualisierungen aufzubereiten.

In den letzten Jahren ist das Interesse von Bibliotheken am Thema (Daten-)Visualisierungen stark gestiegen und auch Einbindungen entsprechender Visualisierungen in Bibliothekskataloge werden häufiger. Dabei werden die Daten und Bestände mittlerweile meist dynamisch visualisiert, so dass Nutzer*innen mit den Daten interagieren und die visualisierten Bestände oder Sammlungen virtuell erkunden können. Auch die DNB nutzt Visualisierungen und hat bereits verschiedene Erfahrungen in diesem Bereich gesammelt: So wurde bereits 2017 in Kooperation mit der FH Potsdam die prototypische Visualisierungsanwendung *DNBVIS*[3] entwickelt, die das virtuelle Browsen und Durchstöbern der Bestände der DNB ermöglichen sollte.[4] Der *GND Explorer*[5] wurde als visuelles Recherchetool für die Gemeinsame Normdatei entwickelt, welches Anwender*innen einen „komfortablen und gleichzeitig umfassenden Zugang zur GND sowie ihrem semantischen Netz bieten"[6] soll. Auch die Beta-Version des neuen Katalogs der DNB[7] enthält einen visualisierten Zeitstrahl als Filter, der nicht nur die zeitliche Verteilung der gefundenen Treffer darstellt, sondern auch zu deren weiterer Eingrenzung genutzt werden kann.

Gleichzeitig entstand der Wunsch, interessierten Nutzer*innen einen ersten Überblick über spezifische Datensets, wie sie z.B. im DNBLab[8] zum Download bereitstehen, zu bieten. Auch für Wissenschaft und Forschung, die sich zunehmend für bibliothekarische Daten interessieren, bietet die Visualisierung eines Datensets Vorteile. So kann eine Visualisierung neben einem ersten Eindruck des zu erwartenden Inhalts beispielsweise auch weitere (Forschungs-)Fragen an die Daten inspirieren. Da sich dynamische Visualisierungen hierfür besonders anbieten, wurde die prototypische und experimentelle Web-basierte Visualisierungsanwendung *DNBVIS_frodiss* entwickelt. Als zugrundeliegendes

1 Erreichbar unter https://dnbvis-frodiss.streamlit.app/, Stand: 18.07.2024.
2 Der Beitrag wurde auf der 112. BiblioCon in Hamburg am 6. Juni 2024 als Teil des Vortragsblocks „Daten im Blick" des Themenkreises 4: „Forschungsnahe Dienste und Open Science" präsentiert.
3 https://dnbvis.fh-potsdam.de, Stand: 18.07.2024.
4 Die Anwendung beinhaltete mit den ihr zugrundeliegenden ca. 15 Millionen Titeldaten aus dem Jahr 2017 bereits bei Veröffentlichung nur einen Teil des Gesamtbestandes der DNB und wurde seither nicht weiter ausgebaut oder aktualisiert.
5 Der GND Explorer ist unter https://explore.gnd.network/ erreichbar, Stand: 18.07.2024.
6 https://explore.gnd.network/, Stand: 18.07.2024.
7 https://katalog.dnb.de/, Stand: 18.07.2024.
8 Erreichbar unter: https://www.dnb.de/dnblab, Stand: 18.07.2024.

Datenset wurde das Set der „Freien Online Hochschulschriften“ der DNB ausgewählt[9], welches die Metadaten der bei der DNB abgelieferten, frei zugänglichen und elektronisch veröffentlichten Dissertationen sowie Habilitationen enthält.

Als experimentelles Projekt standen für diese Entwicklung keine zusätzlichen Ressourcen zur Verfügung, vielmehr erfolgte sie „nebenbei“. Aufgrund dieses Charakters wurden für ähnliche Visualisierungen teilweise abweichende Vorgehensweisen zur Erprobung genutzt. Die Daten wurden so verwendet, wie sie im Set vorgefunden wurden und mit Hilfe existierender Dokumentationen verarbeitet. Weitere Expert*innen für eine mögliche Verbesserung der Methodik, wie etwa hinsichtlich der Inhaltserschließung, wurden daher nicht hinzugezogen.

2. Entwicklung von *DNBVIS_frodiss*

Der Entwicklungsprozess umfasste mehrere aufeinander aufbauende Schritte: Zunächst wurden die Daten aus dem heruntergeladenen Set in einer Jupyter-Notebook-Umgebung mit Python in ein Pandas Dataframe[10] konvertiert und im Anschluss aufbereitet[11]. Die Visualisierungen für die Anwendung wurden ebenfalls in Python erstellt, hierfür wurde vor allem die Bibliothek Plotly genutzt. Die interaktive Webanwendung wurde mit dem Open-Source-Framework Streamlit erstellt, das es ermöglicht, Python-Skripte direkt in interaktive Webanwendungen umzuwandeln. Alle für die Anwendung benötigten Dateien liegen in einem GitHub-Repository[12]. Für das Deployment wird die Anwendung auf Streamlit Cloud gehostet, welches direkt auf das Repository zugreift und bei Aktualisierungen automatisch die Anwendung updatet. Die Extraktion der Daten aus dem alle vier Monate neu zur Verfügung gestellten Dump und die verschiedenen Vorverarbeitungsschritte erfolgen via Skript. Jedoch müssen diese Schritte bislang manuell ausgeführt und die neu erstellten Visualisierungsdaten im Anschluss auf Github abgelegt werden.

2.1 Datengrundlage und Preprocessing

Um die verschiedenen gewünschten Visualisierungen zu erstellen, wurde zunächst festgelegt, welche Informationen aus den MARC21-Datensätzen benötigt werden und daher extrahiert werden sollten. Als Visualisierungen geplant waren eine Übersicht der im Set enthaltenen Hochschulschriften nach Publikationsjahr sowie eine weitere Übersicht nach Fachgebieten. Außerdem sollte eine Kartenansicht der Hochschulschriften nach Publikationsort erstellt werden und eine Übersicht der verschiedenen Sprachen, in denen die Hochschulschriften verfasst wurden. Da der Publikationsort jedoch vom Sitz der jeweils für die Hochschulschrift verantwortlichen Hochschule abweichen kann (z.B. wenn die Hochschulschrift durch einen Verlag publiziert wurde, der nicht am Hochschulort ansässig ist), wurde zudem eine Visualisierung der Hochschulschriften nach Hochschule vorgesehen. Aus den einzelnen

9 Das für *DNBVIS_frodiss* genutzte Datenset steht im DNBLab untezum Download als MARC21-xml bereit und wird 3x jährlich aktualisiert, Stand: 18.07.2024.
10 Hierbei handelt es sich um eine tabellarische Datenstruktur.
11 Für die Datenextraktion und -aufbereitung wurden verschiedenen Python-Bibliotheken genutzt, unter anderem Pandas, lxml, BeautifulSoup und unicode.
12 Erreichbar unter: https://github.com/deutsche-nationalbibliothek/dnbvis_frodiss, Stand: 18.07.2024.

Datensätzen sollten daher neben der jeweiligen ID-Nummer (IDN) des Datensatzes Informationen zu Sprache, DDC, Sachgruppe, Publikationsjahr, Publikationsort sowie der Hochschulschriftenvermerk extrahiert werden.

Hierfür wurde zunächst ein Pythonskript geschrieben, welches jeden Datensatz des Datensets nach diesen Informationen durchsucht. Ausgehend von der jeweiligen IDN des Datensatzes werden diese Informationen in ein Dictionary überführt, welches dann in ein Pandas Dataframe konvertiert wird. Dafür wurden die Inhalte der folgenden MARC21-Felder extrahiert:

- IDN: Controlfield 001
- Sprache: Datafield 041 $a
- DDC/Sachgruppe: Datafields 082 $a, 083 $a, 083 $q und 084 $a
- Publikationsjahr: Datafields 264 $a, 502 $c, Controlfield 008
- Publikationsort: Datafields 264 $a, 502 $a, 502 $c
- Hochschulschriftenvermerk: Datafields 502 $a, 502 $b, 502 $c, 502 $d

Zusätzlich wurden die Felder

- Creator: Datafields 100 $a und 110 $a
- Title: Datafields 245 $a und $b
- Notes: Datafields 500 $a

extrahiert, um Ergebnisse besser prüfen zu können, fehlende Informationen zu ergänzen und später ggf. weitere Visualisierungen hinzufügen zu können.

```
#Creator:
creator = xml.findall("marc:datafield[@tag = '100']/marc:subfield[@code = 'a']", namespaces=ns)
creator_body = xml.findall("marc:datafield[@tag = '110']/marc:subfield[@code = 'a']", namespaces=ns)
if creator:
    creator = creator[0].text
elif creator_body:
    creator = creator_body[0].text
else:
    creator = "fail"
```

Abb. 1: Code-Snippet zur Datenextraktion der Information „Creator“

Bereits bei der Überführung der Informationen aus dem MARC21-xml in ein Dataframe mussten Informationen entsprechend der bibliothekarischen Regelwerke aus verschiedenen Feldern bezogen und priorisiert werden. So wird z.B. nach der Information für die Angabe „Creator“ zunächst im Datafield 100 $a gesucht und bei positivem Ergebnis in eine Variable „creator“ geschrieben. Existiert das Feld nicht, wird in einem zweiten Schritt das Datafield 110 $a geprüft und dessen Inhalt in die Variable übernommen. Wird auch hier nichts gefunden, wird der Inhalt der Variable „creator“ auf „fail“ gesetzt. Der Inhalt der Variable wird dann in die Spalte „Creator“ des Dataframes geschrieben und

unterscheidet an dieser Stelle nicht zwischen Person und Körperschaft.[13] Für andere Informationen funktioniert das Verfahren entsprechend der Regelwerke analog.

Bei einer ersten Sichtprüfung des so erstellten Dataframes fiel auf, dass im Set einige Datensätze enthalten waren, die überraschenderweise nicht über einen Hochschulschriftenvermerk verfügten. Anhand der IDNs konnte schnell festgestellt werden, dass es sich hierbei um sogenannte übergeordnete Datensätze handelt, d.h. meist um Datensätze verschiedener Reihen, in denen Hochschulschriften erscheinen. Der Datensatz selbst ist aber keiner konkreten Hochschulschrift zugeordnet und verfügt daher auch nicht über einen Hochschulschriftenvermerk. Da diese Datensätze für die angestrebte Visualisierung der frei verfügbaren Hochschulschriften keine Relevanz haben, wurden sie aus dem Dataframe entfernt.

2.2 Erstellung der einzelnen Visualisierungen

Für die einzelnen Visualisierungen mussten die Daten jeweils spezifisch vorbereitet werden. Für *DNBVIS_frodiss* wurde dabei für jede Visualisierung eine eigene Datei als Basis erstellt. Statt alle Informationen in einer einzigen Datei zusammenzuführen, aus der sich dann alle Visualisierungen speisen, enthält sie nur die für die Visualisierung notwendigen Informationen. Dieser Ansatz wurde gewählt, um die für jede Visualisierung zu ladende Datenmenge möglichst gering zu halten. Ziel war es, einerseits lange Ladezeiten zu vermeiden und andererseits die RAM-Kapazitäten der durch Streamlit-Cloud zur Verfügung gestellten Hosting-Lösung nicht zu überschreiten. Im Folgenden werden die Schritte für die verschiedenen Visualisierungen kurz vorgestellt:

2.2.1 Publikationsjahre

Die Visualisierung für die Publikationsjahre erforderte nur geringe Anpassungen der Daten. Da im Datenfeld 264 $c neben der einfachen vierstelligen Jahreszahl auch Angaben wie „[2009]“, [2009?], 2009-, ©2009 und andere Varianten vorkommen können, mussten hier zunächst alle nicht-numerischen Zeichen entfernt werden. Außerdem wurde die Darstellung auf die Jahre ab 1990 beschränkt. Dies war möglich, da es sich bei den im Datenset vor 1990 verzeichneten freien, digitalen Hochschulschriften nur um einige wenige, typsicherweise retrodigitalisierte Hochschulschriften handelt.

2.2.2 Fachgebiete

Die Aufbereitung der Daten für die Darstellung der verschiedenen Fachgebiete war aufwändig: Bei Vorhandensein mehrerer Einträge von DDC-Klassen wurden diese bereits während der Datenextraktion nach Ursprung des Datenfelds priorisiert, zusätzlich aber alle Einträge auch in separate Spalten übernommen, um mögliche Fehler im Skript leichter feststellen zu können. Außerdem wurde die Spalte „Sachgruppe“ erstellt, die die früher verwendeten DNB-Sachgruppen enthält und deren Inhalte bei Bedarf zunächst noch auf eine DDC-Klasse gemappt werden mussten.

13 Auf die Abfrage weiterer Felder wurde verzichtet, da die im Set enthaltenen Hochschulschriften per definitionem von einer alleinig verantwortlichen Person verfasst sein müssen, so dass bereits die Suche nach einem Körperschaftsnamen hier theoretisch überflüssig ist und nur als „failsafe“ eingebaut wurde.

Etwa 20.000 Datensätze des knapp 300.000 Datensätze umfassenden ursprünglichen Datendumps des Sets „Freie Online Hochschulschriften“ verfügten nicht über einen Eintrag einer DDC-Klasse. Ein Großteil dieser Datensätze enthielt jedoch einen Sachgruppeneintrag. Mit Hilfe eines Mappings der verschiedenen DNB-Sachgruppen auf die tausend Klassen der dritten Ebene der DDC wurde in einem zweiten Dataframe für diese Datensätze die entsprechende DDC-Klasse automatisiert ermittelt. Beide Dataframes wurden anschließend gemerged. Es blieben dabei knapp 900 Datensätze ohne ermittelbare DDC, die für die weitere Darstellung der Fachgebiete nach DDC nicht berücksichtigt werden konnten.

Auch gab es einen geringen Anteil an Datensätzen, die im entsprechenden MARC21-Datenfeld statt der DDC-Notation deren Klassenbenennung ausgeschrieben enthielten. Da es sich hier um eine niedrige dreistellige Anzahl an Datensätzen handelte, wurde das für die Darstellung der Fachgebiete erstellte Dataframe in Open Refine exportiert und die Benennungen manuell in die zugehörigen Notationen gewandelt. Seit der ursprünglichen Entwicklung der Anwendung *DNBVIS_frodiss* wurde die Datenqualität in Bezug auf die DDC deutlich verbessert, so dass bei Updates der Schritt des Mappings mittlerweile entfallen kann. Ein kleiner Grundstock an Datensätzen ohne jegliche fachliche Zuordnung bleibt jedoch weiterhin bestehen.

Die Visualisierung der Fachgebiete wurde dann durch mehrere Ansätze umgesetzt: Zum einen wurde eine simple Zählung des Vorkommens der 10 DDC-Hauptklassen durchgeführt, in dem jeweils die Häufigkeit der ersten Ziffer der DDC-Notation ermittelt wurde. Diese Angaben wurden für die Kacheldarstellung als erste Übersicht genutzt. Erweitert wurde die Darstellung um die Angabe der im Datenset enthaltenen Datensätze ohne fachliche Zuordnung.

Abb. 2: Übersicht Fachgebiete, Stand der Daten: 28.03.2024

Die so ermittelten Werte wurden zusätzlich für eine Übersicht als Balkendiagramm verwendet. Für eine detaillierte Betrachtung der einzelnen Fachgebiete wurde dann eine per Dropdown-Menü anwählbare Darstellung als Plotly Sunburst Chart gewählt, die es nach Auswahl einer DDC-Hauptklasse erlaubt, die Klassen der zweiten sowie der dritten Ebene genauer zu betrachten und in diese „zu zoomen“:

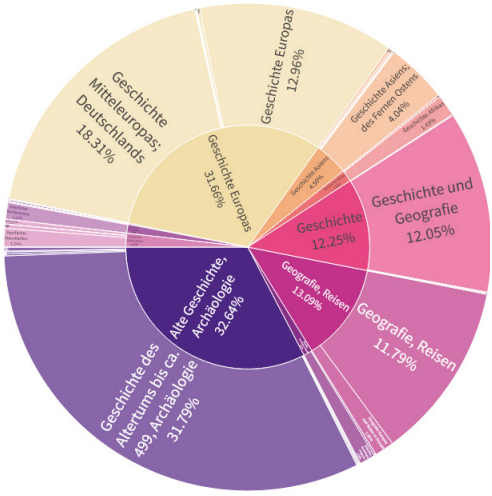

Abb. 3: 900 - Geschichte und Geografie, Stand der Daten: 28.03.2024

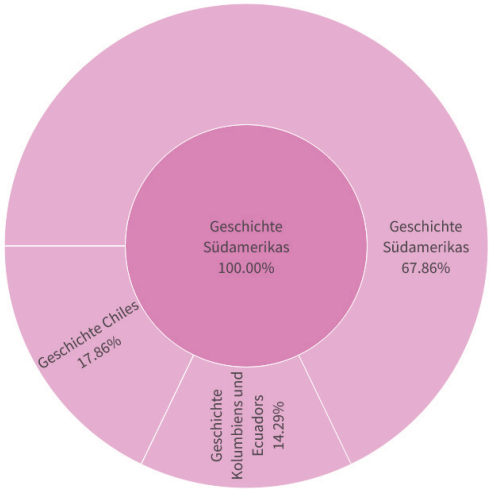

Abb. 4: Zoom auf Geschichte Südamerikas, Stand der Daten: 28.03.2024

2.2.3 Publikationsorte

Für die Darstellung der Publikationsorte und der Hochschulen war die Georeferenzierung der entscheidende Faktor. Um beide miteinander vergleichen und evaluieren zu können, wurden für beide Visualisierungen leicht unterschiedliche Ansätze verfolgt.

Nach der initialen Datenbereinigung lagen die Informationen für die Publikationsorte bereits grob bereinigt vor, so dass im ersten Schritt lediglich ein neues Dataframe erzeugt wurde, welches zunächst nur die IDN und den ermittelten Publikationsort enthielt. Dopplungen des Publikationsortes wurden entfernt und das so entstandene Dataframe als csv exportiert. Mit Hilfe von Open Refine und der Lobid-Reconciliation-API[14] wurde nun ein Abgleich der Ortsnamen mit der GND durchgeführt. Die Tabelle wurde daraufhin um die jeweils ermittelte Schreibweise des Ortes in der GND sowie die zugehörige GND-ID erweitert und wieder als neues Dataframe in die JupyterLab-Umgebung geladen.[15]

14 https://lobid.org/gnd/reconcile, Stand: 18.07.2024.

15 So wurde z.B. die Bezeichnung „Freiburg (Breisgau)“ im Datensatz dem GND-Datensatz „Freiburg im Breisgau“ mit der GND-ID 4018272-1 zugeordnet.

Die so ermittelten GND-IDs wurden nun in eine Liste überführt, über die SRU-Schnittstelle der DNB für die GND abgefragt[16] und die zugehörigen Normdatensätze identifiziert. Aus diesem Set wurden die Geokoordinaten der jeweiligen Ortschaften extrahiert und in ein Dataframe überführt, welches die GND-ID und die ermittelten Werte für Longitude und Latitude enthält. Zuletzt wurden die verschiedenen erstellten Dataframes so zusammengeführt, dass das ursprüngliche Dataframe mit allen Einträgen um die zusätzlich ermittelten Informationen angereichert und daraufhin die Häufigkeit der Ortsnamen ermittelt und nach diesen gruppiert wurde.

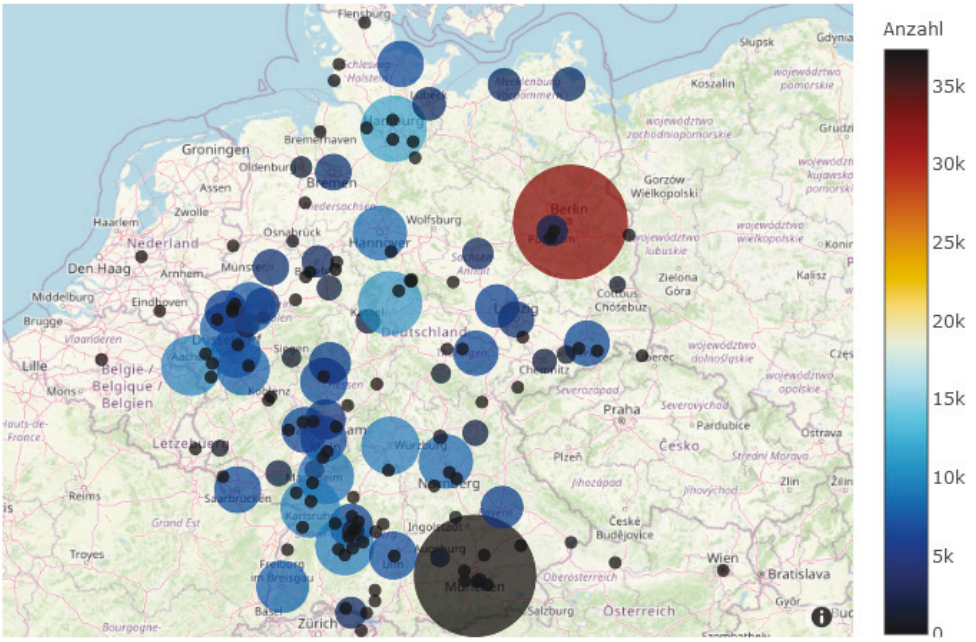

Abb. 5: Darstellung der Hochschulschriften nach Publikationsorten, Stand der Daten: 28.03.2024

2.2.4 Hochschulen

Für die Erstellung der Visualisierung der zu den Publikationen gehörenden Hochschulen wurde ein neues Dataframe erstellt, welches nur aus der IDN und dem Hochschulschriftenvermerk bestand. Der Hochschulschriftenvermerk wurde dann anhand der Kommata in die Spalten „place“, „uni“, und „remaining“ gesplittet.

16 Die SRU-Schnittstelle der DNB ist unter https://services.dnb.de/sru erreichbar, der Katalog der Gemeinsamen Normdatei unter https://services.dnb.de/sru/authorities. Die Dokumentation der SRU-Stelle kann unter https://www.dnb.de/sru eingesehen werden. Stand: 18.07.2024.

	ID	diss	place	uni	remaining
0	1000000389	freiburg (breisgau), univ., diss., 2009	freiburg (breisgau)	univ.	diss., 2009
1	1000006395	düsseldorf, univ., diss., 2009	düsseldorf	univ.	diss., 2009
2	100000645X	bayreuth, univ., diss., 2009	bayreuth	univ.	diss., 2009
3	1000006948	bayreuth, univ., diss., 2010	bayreuth	univ.	diss., 2010
4	1000007138	duisburg, essen, univ., diss., 2009	duisburg	essen	univ., diss., 2009
...	...	...	...	...	...
324777	999997297	göttingen, univ., diss., 2009	göttingen	univ.	diss., 2009
324778	999997505	bayreuth, univ., diss., 2009	bayreuth	univ.	diss., 2009
324779	99999767X	bayreuth, univ., diss., 2009	bayreuth	univ.	diss., 2009
324780	999999389	freiburg (breisgau), univ., diss., 2009	freiburg (breisgau)	univ.	diss., 2009
324781	999999540	freiburg (breisgau), univ., diss., 2009	freiburg (breisgau)	univ.	diss., 2009

324782 rows × 5 columns

Abb. 6: Ansicht des Dataframes zur Erstellung der Visualisierung nach Publikationen per Hochschule

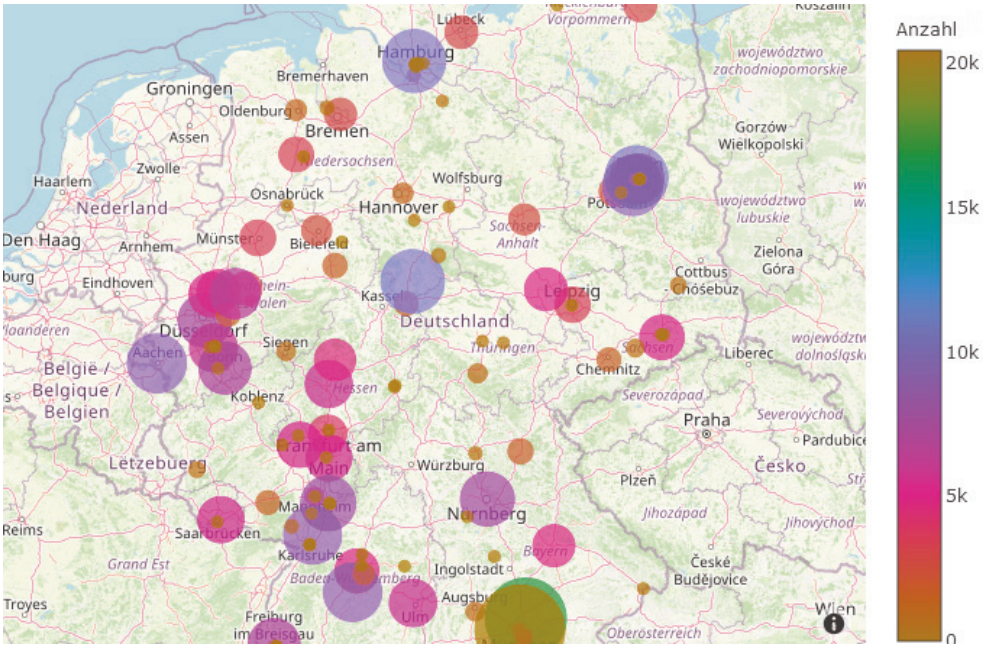

Abb. 7: Darstellung der Hochschulschriften nach Hochschule, Stand der Daten: 28.03.2024

Im Anschluss wurde das Dataframe um die Spalte „Hochschule“ ergänzt, die den offiziellen Namen der jeweiligen Hochschule enthalten soll. Dieser musste zunächst aus den Inhalten des nicht normierten Hochschulschriftenvermerks ermittelt werden, zum Abgleich wurde eine Liste der Hochschulen Deutschlands der Hochschulrektorenkonferenz genutzt.[17] Ein direktes Merging erwies sich hier nicht als zielführend, da Schreibweisen, Abkürzungen und Bezeichnungen der Hochschulen im Hochschulschriftenvermerk zu divers und nicht immer eindeutig sind. Stattdessen wurde ein Skript erstellt, welches nach verschiedenen Kombinationen von Strings sucht und diesen die entsprechende Hochschule zuordnet. Besonders herausfordernd war die Zuordnung in Städten mit mehreren Hochschulen, wie z.B. Darmstadt, wo Begriffe wie „Uni“, „Hoch“, „Tech“, „TU“ und „TH“ sowie Variationen genutzt werden.[18] Bei einigen Hochschulen wie in Berlin bleibt eine gewisse Ungenauigkeit, da die Abkürzung „Uni“ im Hochschulschriftenvermerk hier sowohl für die „Freie Universität Berlin“, die „Humboldt-Universität zu Berlin“ als auch für die „Technische Universität Berlin“ stehen kann und sich nicht eindeutig auflösen lässt, ohne die jeweilige Hochschulschrift zu konsultieren. Hier wurde mit dem Textzusatz „ – likely“ zwar ein Marker gesetzt, um die fehlende Eindeutigkeit sichtbar zu machen, in der Kartendarstellung der Hochschulen werden diese Publikationen aktuell aber der FU Berlin zugeordnet. Eine Verbesserung des Algorithmus ist hier wünschenswert und wird, die Ressourcen vorausgesetzt, angestrebt.

Im Anschluss wurden die Einträge für die einzelnen Hochschulen gruppiert und gezählt. Da sich die Darstellung der Hochschulen im Gegensatz zur Darstellung der Publikationsorte durch das oben beschriebene Vorgehen auf Deutschland bezieht, wurden einige Publikationen ausgeschlossen und es bleiben 135 verschiedenen Hochschulen. Die ermittelten offiziellen Namen der Hochschulen wurden im Anschluss zur Georeferenzierung mit der Bibliothek Geopy genutzt. Dieser Prozess schlug bei 16 Hochschulen fehl, für die auf das bereits bei den Publikationsorten genutzte Verfahren zurückgegriffen wurde. Nach Ermittlung der GND-ID wurden mit Hilfe mehrerer SRU-Abfragen Latitude und Longitude aus der GND ermittelt. Zum Schluss wurden die Daten in einem Dataframe zusammengeführt, welches der Visualisierung zugrunde liegt.

	lang	counts	name
0	aar	10	Danakil-Sprache
1	ara	3	Arabisch
2	ben	2	Bengali
3	bos	1	Bosnisch
4	chi	1	Chinesisch
5	dan	1	Dänisch
6	dut	2	Niederländisch
7	eng	134079	Englisch
8	enm	1	Mittelenglisch
9	fail	5	Ohne Sprache
10	fre	180	Französisch
11	ger	189750	Deutsch
12	grc	1	Griechisch (Altgriechisch)

Abb. 8: Ausschnitt des Dataframes für die Erstellung der Sprachencode-visualisierung

2.2.5 Sprachen

Für die Erstellung der Visualisierung der Publikationssprache wurde ein neues Dataframe erstellt, welches nur den Sprachcode enthielt. Die Codes wurden gruppiert und gezählt. Zusätzlich wurde eine Tabelle importiert, die die Auflösung der

17 Liste der im von der Hochschulrektorenkonferenz angebotenen Portal Hochschulkompass gemeldeten Hochschulen: https://de.wikipedia.org/wiki/Liste_der_Hochschulen_in_Deutschland, Stand: 18.07.2024.

18 Die Abkürzungen wurden der „Technischen Universität Darmstadt“ bzw. der „Hochschule Darmstadt“ zugeordnet.

Sprachencodes in deren Bezeichnungen nach ISO 639-2/B enthielt. Das neue Dataframe wurde um diese Bezeichnungen erweitert.

Da die meisten im Datenset enthaltenen Publikationen auf Deutsch oder Englisch vorliegen und andere Sprachen nur einen sehr geringen Anteil ausmachen, wurden zwei verschiedene Kreisdiagramme zur Visualisierung erstellt. Das erste Diagramm zeigt die Sprachenverteilung des gesamten Datensets und unterscheidet zwischen Publikationen auf Deutsch, Englisch oder anderen Sprachen. Das zweite Diagramm zeigt dann die prozentualen Anteile der unter „andere Sprachen" gruppierten Sprachen:

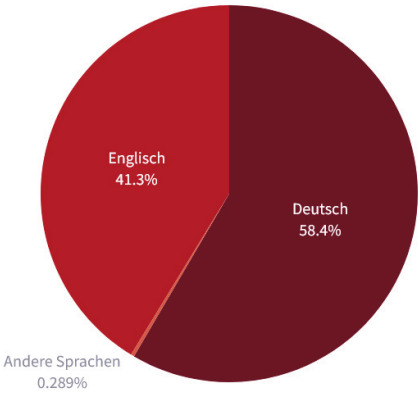

Abb. 9: Sprachen: Gesamtansicht, Stand der Daten: 28.03.2024

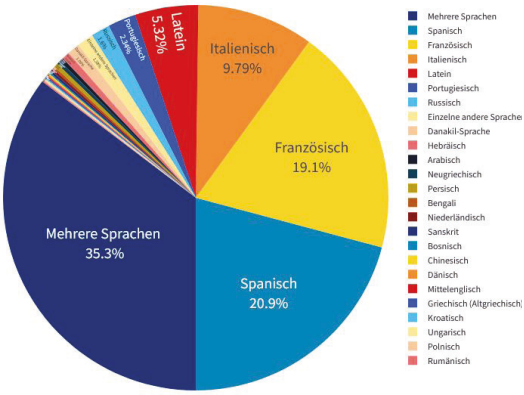

Abb. 10: Alle in Abb. 6 unter „Andere Sprachen" enthaltenen Sprachen, Stand der Daten: 28.03.2024

3. Fazit

Die Entwicklung von *DNBVIS_frodiss* barg einige Herausforderungen, zeigte aber auch verschiedene Potentiale und sinnvolle Weiterentwicklungsmöglichkeiten auf. Auf einige dieser Herausforderungen soll im Folgenden kurz eingegangen werden:

Freitextfelder bedürfen häufig besonders komplexer Lösungen, müssen die Daten hier doch individuell betrachtet und zunächst auf verschiedenen Wegen aufbereitet werden, um für Visualisierungen nutzbar zu sein. So stellen bei der Erfassung im Hochschulschriftenvermerk verschiedene Schreibweisen, Abkürzungen, Namensänderung von Institutionen sowie simple Tippfehler die größte Herausforderung dar. Eine intellektuelle Prüfung der bereinigten Daten ist hier bislang unumgänglich, birgt die rein algorithmische Bereinigung doch durch die Diversität und die Mehrdeutigkeit der Daten ein zu großes Fehlerpotential. Je größer die Datenmenge, desto schwieriger gestaltet sich eine manuelle Qualitätskontrolle. Die entworfenen Algorithmen zur Bereinigung müssen mit Hinzukommen neuer Datensätze beim Update des Datasets häufig erweitert werden, um neu auftretende Einzelfälle zu berücksichtigen.

Auch birgt die Georeferenzierung verschiedene Tücken. Zwar bieten Spezialbibliotheken hier etwas Abhilfe, doch treten schnell Probleme auf, wenn etwa Städtenamen nicht erkannt oder durch Mehrdeutigkeiten falsche Geokoordinaten zugeordnet werden. Auch die Nutzung der GND in diesem Kontext bedarf eines gewissen intellektuellen und manuellen Aufwands, denn auch hier führen Mehrdeutigkeiten dazu, dass Geografika teilweise händisch bestätigt werden müssen und mehrere Abfragen und Datenextraktionen notwendig sind, um die passenden Geokoordinaten aus den als relevant ermittelten GND-Datensätzen zu extrahieren.

Nicht zuletzt ist auch die wachsende Größe des Datensets „Freie Online Hochschulschriften" eine Herausforderung: Alle hier beschriebenen Verarbeitungsschritte werden aktuell lediglich auf einem handelsüblichen Arbeits-PC durchgeführt. Dabei hat das Datenset „Freie Online Hochschulschriften" in den letzten zwei Jahren durch den Zuwachs an Publikationen inzwischen eine Größe erreicht, für die die RAM-Kapazitäten des Rechners zur Verarbeitung nicht mehr ausreichend sind. Neben der Lösung, auf einen leistungsstärkeren Rechner auszuweichen, bietet sich hier insbesondere eine Verbesserung der Effizienz der Skripte durch eine Umstellung von Pandas hin zu Polars[19] an. Erste Tests haben diese Vermutung eindrucksvoll bestätigt, jedoch würde die Umstellung eine Anpassung aller Skripte und aller darin enthaltenen Verarbeitungsschritte voraussetzen, welche aus Ressourcengründen bislang noch nicht erfolgen konnte.

DNBVIS_frodiss erfüllt das Ziel, mit Hilfe einer Visualisierungsanwendung einen schnellen Über- und Einblick in ein Datenset bieten zu können. Gleichzeitig verdeutlicht die Entwicklung der Anwendung nicht nur deren Potentiale, sondern regt auch zur Weiterentwicklung an. Das Konzept eignet sich insbesondere, um neue Zugänge zu großen, anderweitig wenig überschaubaren Datenmengen zu bieten. Es kann problemlos um weitere Zugänge erweitert sowie auf andere Datensets angewendet werden. Auch eignet sich diese Form der Datenaufbereitung durchaus, um neue Denkanstöße zur Sammlungsentwicklung, zu internen Workflows, zur Qualitätssicherung oder zu verschiedenen Forschungsfragen zu geben.

Zur Weiterentwicklung bietet sich neben Verfeinerungen der Methodik zur Steigerung der Präzision und Effizienz bei der Datenverarbeitung die Einbindung weiterer Visualisierungen wie bspw. einer Wordcloud an. Eine Anpassung der Skripte von Pandas auf Polars wäre ebenso wie ein höherer Grad an Automatisierung der jeweiligen Verarbeitungsschritte wünschenswert, um die Anwendung dauerhaft und mit möglichst wenig Ressourcenaufwand betreiben zu können. Darüber hinaus denkbar ist die Entwicklung weiterer Tools wie z.B. einer Anwendung, mit der Nutzer*innen für beliebige MARC21-xml-Metadatensets die entsprechenden Visualisierungen automatisiert erstellen und nachnutzen können.

Stephanie Nitsche, Deutsche Nationalbibliothek Frankfurt, https://orcid.org/0000-0003-1624-896X

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/6084

19 Mehr Informationen zu Polars: https://pola.rs/, Stand: 18.07.2024.