Ramon Voges, André Wendler

Ein Dashboard für die GND

Wie kann man eine Datenbank mit 9 Millionen Einträgen erfassen und überblicken? Wie kann man sie analysieren, Fehler finden und verbessern? Und wie kann man das gespeicherte Wissen möglichst vielen Personen zugänglich machen?

Diese und ähnliche Fragen waren der Ausgangspunkt für die Arbeit am GND-Dashboard, das unter der Internetadresse <https://share.streamlit.io/deutsche-nationalbibliothek/gnd-dashboard/main/dashboard/gnd-app.py> frei zugänglich ist. Die Gemeinsame Normdatei (GND) enthält normierte Bezeichnungen für Personen, Körperschaften, Geografika, Sachschlagworte und weitere Normdaten. Sie stammt ursprünglich aus der Bibliothekswelt und wird in der Deutschen Nationalbibliothek (DNB) gehostet. Kultureinrichtungen, die Normbegriffe der GND verwenden, können mit ihrer Hilfe Namen, Institutionen und geografische Orte eindeutig identifizieren.

Seit einigen Jahren nutzen neben Bibliotheken zunehmend auch Museen, Archive und andere Forschungseinrichtungen diese Normdaten. Dadurch wird es immer leichter, die Datenbestände und Sammlungen der einzelnen Einrichtungen miteinander zu verbinden.

Sichtbarkeit

Anlässlich der ersten GND Convention, GNDcon, die Ende 2018 in den Räumen der DNB in Frankfurt am Main physisch und per Stream auch in Leipzig stattfand und die die Öffnung der GND für andere Kultur- und Forschungseinrichtungen voranbringen sollte, hat das Deutsche Buch- und Schriftmuseum der Deutschen Nationalbibliothek (DBSM) eine antike Steintrommel mit einem Himmel aus Normdaten versehen.[1] Das Exponat machte auf diese Weise anschaulich, wie materielle Objekte – die bisweilen mehrere hundert Kilogramm schwer sein können – durch luftige Daten in einen historischen Zusammenhang gestellt und damit zum Sprechen gebracht werden können.

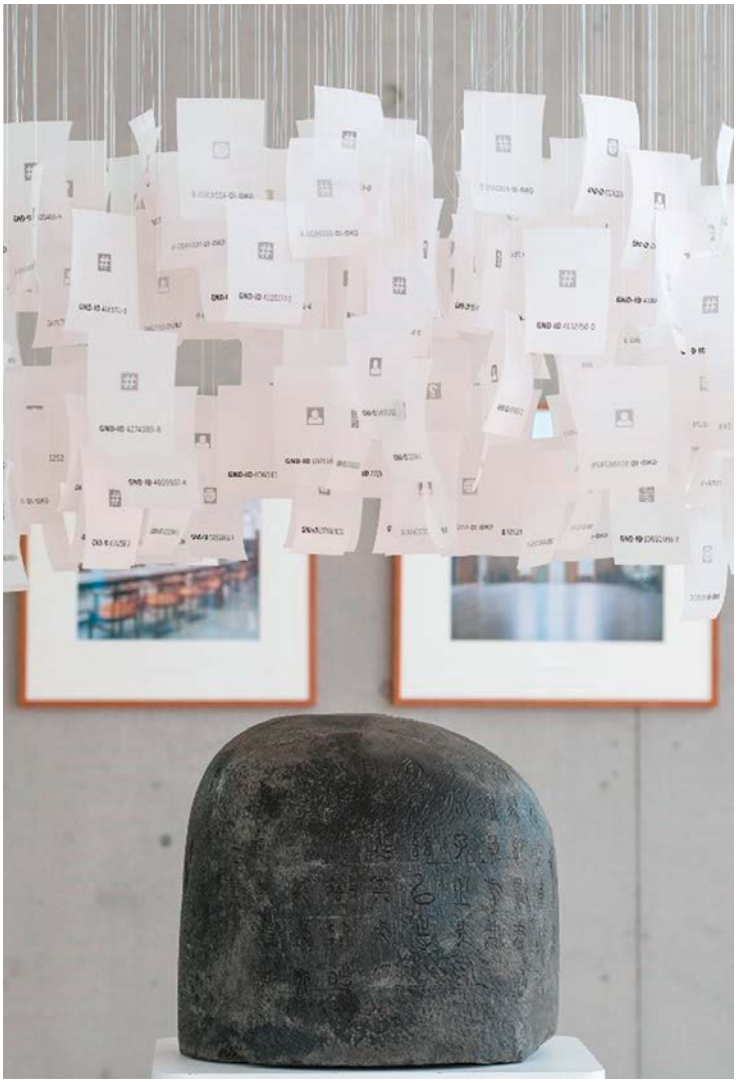

Abb. 1: Chinesische Steintrommel unter einer Wolke aus GND-Daten. Foto: Deutsche Nationalbibliothek, Stephan Jockel.

Für die GNDcon 2.0, die wegen der Covid-19-Pandemie vom 7. bis 11. Juni 2021 als rein virtuelle Veranstaltung stattfinden musste,[2] wäre ein materielles Exponat unpassend gewesen. Auf Anregung der Veranstalter*innen berieten deswegen die Teilnehmer*innen des Python Meetups der DNB, wie ein digitales Exponat für die GNDcon beschaffen sein könnte. Rasch war die Idee eines interaktiven Dashboards geboren, das die enorme Fülle an Informationen, die in der GND gespeichert sind, anschaulich aufbereiten sollte.

Abb. 1: Screenshot des GND-Dashboards.
Abrufbar unter: https://share.streamlit.io/deutsche-nationalbibliothek/gnd-dashboard/main/dashboard/gnd-app.py

Big Data

Wenn Mitarbeitende in Museen oder Bibliotheken mit GND-Sätzen arbeiten, suchen sie üblicherweise nach bestimmten Personen, Sachbegriffen oder Institutionen. Sie ergänzen diese vielleicht um neue Informationen oder verknüpfen den Datensatz mit einem anderen, etwa eine Person mit einer Institution, der sie einmal angehörte. Auch Gruppen von Datensätzen können einem begegnen, wie zum Beispiel alle Angehörigen einer bestimmten Institution. Die GND als ganzes Datenset dürfte allerdings seltener in den Blick geraten.

Wie schwierig es ist, die ganze GND zu überblicken, zeigen schon wenige Eckdaten. Im Juli 2021 befanden sich 8,95 Mio. Entitäten in der GND. Der größte Teil davon waren knapp 5,6 Mio. Personensätze (Tp), gefolgt von ca. 1,5 Mio. Körperschaften (Tb). Ein Gesamtabzug der Daten im Ursprungsformat Pica+ ist derzeit über 5 Gigabyte groß. Dateien dieser Größe kann man nicht einfach öffnen und durchsuchen. Für solche Datenmengen sind besondere Methoden erforderlich, wie sie in den Data Sciences entwickelt werden.

Um die Daten für das GND-Dashboard aufzubereiten, kam ein spezielles Programm zum Einsatz, ein sogenannter Parser. Ein solches Programm, das auf die Verarbeitung von Pica+-Daten spezialisiert ist, hat ein Kollege aus dem Referat »Automatische Erschließungsverfahren und Netzpublikationen« der DNB geschrieben.[3] Mit diesem Tool lässt sich der Gesamtabzug der GND in wenigen Minuten

verarbeiten. Dabei können einzelne Merkmale ausgefiltert, Daten extrahiert und einfache Statistiken erstellt werden. Als Ergebnis liefert das Tool verhältnismäßig kleine Textdateien im CSV-Format, die sehr viel unkomplizierter weiterverarbeitet werden können als der Gesamtabzug.

Nationalökonomie und interaktive Dashboards

In der Covid-19-Pandemie wurden Dashboards mit Datenanalysen zu Pfeilern der öffentlichen Debatte über die epidemiologischen Maßnahmen. Sie zeigten, wie sich weitverteilte und unübersichtliche Ereignisse als Datendarstellungen erfassen lassen. Bereits seit dem 18. Jahrhundert finden grafische Verfahren zur Darstellung großer numerischer Zusammenhänge Verwendung. Der schottische Ingenieur und Wirtschaftswissenschaftler William Playfair gilt als einer der Erfinder von Balken- und Kreisdiagrammen. Er verwendete sie, um die volkswirtschaftlichen Zusammenhänge seiner Zeit darzustellen. Bis ins 20. Jahrhundert waren Infografiken kuratierte Darstellungen, die von darauf spezialisierten Infografiker*innen für Atlanten, statistische Jahrbücher oder Zeitungsveröffentlichungen produziert wurden.[4]

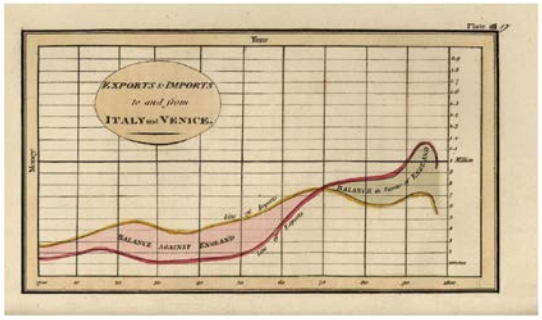

Abb. 2: Der Schotte William Playfair nutzte Infografiken, um volkswirtschaftliche Zusammenhänge darzustellen. Hier vergleicht er Ex- und Importe von Venedig und Italien

Mit dem Aufkommen der digitalen Datenverarbeitung seit den 1940er-Jahren und insbesondere dem »Personal Computer« seit den 1970er-Jahren konnten statistische Darstellungen dann quasi überall mit relativ wenig Aufwand produziert werden. Parallel dazu wurden immer mehr Daten produziert und erhoben. Mittlerweile ist das Bild der Daten als Öl des 21. Jahrhunderts schon etwas abgegriffen. Es beschreibt aber treffend den Umstand, dass Datensätze heute nicht mehr erhoben, abgeschlossen und dann veröffentlicht werden, sondern dass wir uns in vielen Kontexten an die nahezu gleichzeitige Erhebung, Auswertung und Darstellung von Datenströmen, sogenannte Echtzeitdaten, gewöhnt haben. Ein Blick auf die eigene Smartwatch zeigt jederzeit, wie weit man vom täglichen Fitnessziel noch entfernt ist. Börsenkurse stehen nicht mehr am nächsten Tag in der Zeitung, sondern sind in Apps und auf Webportalen jederzeit verfügbar. Wer wissen will, ob man den Gang zum Bäcker trockenen Fußes wagen kann, schaut seltener zum Himmel als auf das Regenradar einer Wetter-App.
Die Rechenkapazitäten selbst tragbarer Rechner sind heute so immens, dass auch große Datensammlungen interaktiv bereitgestellt werden und die Nutzer*innen durch Filter und Facettierung ihre ganz eigenen Fragen an die Daten richten können. Dazu gehören allerdings grundlegende Kenntnisse in Statistik und Data Literacy, die noch nicht überall Teil der allgemeinen Bildung geworden sind.

GND interaktiv

Für die GND bietet sich diese interaktive Art der Analyse an, weil sehr unterschiedliche Institutionen und Personengruppen ganz verschiedene Fragen an den Datenbestand haben. Das Dashboard zeigt zunächst einfache, quantitative Daten zur Anzahl einzelner Satzarten und der Verteilung der Katalogisierungslevel innerhalb der Satzarten. Die am häufigsten vorkommenden Relationierungscodes werden dargestellt sowie die Zahl der monatlich neu angelegten Datensätze seit 1972. Die häufigsten Sachbegriffe der jüngsten zehn Tage der Datenbasis erscheinen in Wordclouds.

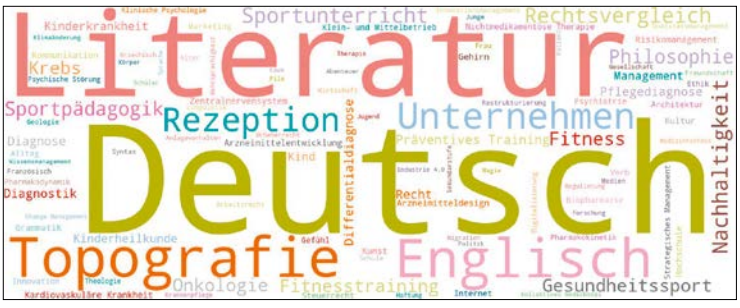

Abb. 3: Wordcloud aus dem Dashboard, 6. August 2021

Darüber hinaus gibt es einige Spezialauswertungen, die die Möglichkeiten aggregierter Normdaten etwas tiefer ausschöpfen. So wurden zum Beispiel die erfassten Wirkungsorte aller GND-Personen ausgewertet. Eine andere Analyse zeigt die Wirkungszentren der Musik zwischen 1400 und 2010, wie sie sich aus einer Kreuzanalyse der im Deutschen Musikarchiv erfassten Musikalien und den damit verbundenen Normdaten ergibt.

Desiderate und Pläne

Eine Auswertung zeigt die am häufigsten in DNB-Titeldaten verlinkten Personen. Unter den Top 10 befindet sich keine Frau. Das ist einerseits ein Abbild der überlieferten, männlich dominierten Kultur, muss aber wohl auch als Mahnung an alle Einrichtungen stehen, die Daten in der GND erfassen, auf die Sichtbarkeit von Frauen achtzugeben. Von allen Personen, deren Geschlecht in der GND erfasst wurde, sind nur 28 Prozent als weiblich markiert, 72 Prozent dagegen als männlich. Solche Zahlen deuten auf eine grobe Missrepräsentation von Frauen in der GND hin, der wir uns stellen sollten. In der nächsten Version des Dashboards werden die Top 10 der verlinkten Personen deshalb getrennt nach Geschlechtern dargestellt.

Sämtliche Skripte und Daten des Dashboards sind unter einer offenen Lizenz frei auf dem Code-Portal GitHub verfügbar. Dort können sich alle Nutzer*innen mit Vorschlägen, eigenen Widgets oder Code-Verbesserungen einbringen.[5] Die Daten werden auf absehbare Zeit monatlich aktualisiert.

Anmerkungen

1 Vgl. dazu auch den Beitrag »GNDCon 2018« von Barbara Fischer und Jürgen Kett im »Dialog mit Bibliotheken« 2019, H. 1, S. 51–53, <https://nbn-resolving.org/urn:nbn:de:101-2019022844>

2 Vgl. <https://wiki.dnb.de/display/GNDCON>

3 Vgl. <https://github.com/deutsche-nationalbibliothek/pica-rs>

4 William Playfair: The commercial and political atlas representing by means of stained copperplate charts, the progress of the commerce, revenues, expenditure, and debts of england, during the whole of the eighteenth century. London 1801, Tafel 17.

5 <https://github.com/deutsche-nationalbibliothek/gnd-dashboard>

