Liebe und Tod in der Deutschen Nationalbibliothek Der DNB-Katalog als Forschungsobjekt der digitalen Literaturwissenschaft

Fischer, Frank
ffischer@hse.ru
Higher School of Economics, Moskau

Jäschke, Robert
r.jaschke@sheffield.ac.uk
Humboldt-Universität, Berlin

Einleitung

Der Sammelauftrag der Deutschen Nationalbibliothek (DNB) beginnt 1913 und bezieht sich auf »lückenlos alle deutschen und deutschsprachigen Publikationen« (»Wir über uns«, 16.03.2017). Der DNB-Katalog ist natürlich längst digitalisiert und die Arbeit mit ihm mittlerweile sehr komfortabel, da der Datendienst der DNB unter http://www.dnb.de/datendienst vierteljährlich einen Komplettabzug der Katalogdaten im RDF-Format bereitstellt, unter der freien Lizenz CC0 1.0. Momentan (Stand vom 23.06.2017) enthält er 14 102 309 Datensätze, also Metadaten zu von der DNB gesammelten Medien. Bisher gibt es aus geisteswissenschaftlicher Sicht nur wenige Versuche, diese Quelle nutzbar zu machen (eine Ausnahme bilden etwa Häntzschel u. a. 2009). Wir präsentieren ein einfaches Framework, mit dem verschiedene Aspekte des DNB-Katalogs untersucht werden können, seine Entwicklung über die knapp 105 Jahre seit Bestehen der Nationalbibliothek (vgl. auch Schmidt 2017, der für die Library of Congress einen ähnlichen Ansatz vorgestellt hat). Wir konzentrieren uns dabei auf Romane als Untersuchungsobjekt, von denen in der DNB rund 180 000 als solche rubriziert sind (dies entspricht nicht der Gesamtanzahl an Romanen, denn Nachauflagen und Übersetzungen zählen dort mit hinein – außerdem fehlen auch einige Romane, da sie nicht entsprechend verschlagwortet worden sind. Dieser Vortrag ist methoden-, nicht vorderhand ergebniszentriert, wobei wir an zwei Anwendungsszenarien aus der Praxis der digitalen Literaturwissenschaft demonstrieren, wie Katalogmetadaten bei der Bearbeitung konkreter Forschungsfragen behilflich sein können bzw. diese überhaupt erst ermöglichen.

Beschreibung des Frameworks

Die Titeldaten der DNB werden in typischen Linked-Data-Formaten (RDF/XML, JSON-LD usw.) angeboten. Der übliche Ansatz mit solchen Daten zu arbeiten ist, diese in eine geeignete Datenbank (Triple-Store) einzuladen und Anfragen mit Hilfe der entsprechenden Anfragesprache (i. A. SPARQL) zu stellen. Prinzipiell sind auch andere Systeme (z. B. relationale Datenbank, Suchmaschine) geeignet. Dies ermöglicht sehr flexible Anfragen und die leichte Einbindung weiterer Datenquellen. Da die Größe der Daten (unkomprimiert ca. 21 GB) jedoch gewisse Anforderungen an die Hardware stellt und die Konfiguration und Optimierung der Datenbank aufwendig ist, haben wir uns für eine andere, kompakte und leichter nachzuvollziehende Lösung entschieden. Langfristiges Ziel ist jedoch die Bereitstellung einer fertig konfigurierten Arbeitsumgebung in Form eines Docker-Containers, in der die Daten in einer Datenbank ad hoc verfüg- und analysierbar sind.

Der Titeldatensatz ist mit 14 102 309 Datensätzen und 227 212 707 Tripeln (»Fakten«) sehr umfangreich und enthält neben Angaben zu Büchern auch Angaben zu weiteren Medientypen wie etwa Zeitschriften. Neben den üblichen Metadatenfeldern wie Titel und Erscheinungsjahr ist bei Buchobjekten meist auch die Seitenanzahl sowie das Format vermerkt. Ganz im Sinne von Linked Data werden viele Angaben mit Hilfe von standardisierten Vokabularien (z. B. Dublin Core oder Bibo) beschrieben und ermöglichen so die Verlinkung mit weiteren Datensätzen. Insbesondere ermöglicht die Angabe der Autor*innen durch die numerische Kennung aus der Gemeinsamen Normdatei (GND) die Verknüpfung der Daten mit Wikidata, der (zukünftig) hinter Wikipedia stehenden Faktendatenbank. Wikidata verwendet ein auf Linked Data basierendes Datenmodell und ermöglicht, ähnlich wie Wikipedia, jedermann das Hinzufügen und Bearbeiten von Daten. Neben Angaben zu Städten und Ländern (z. B. Fläche, Einwohnerzahl) sind in Wikidata auch Daten zu zahlreichen Persönlichkeiten gespeichert, etwa deren Namen, Geburtsdaten, Berufe, Werke und, falls vorhanden, GND-Kennung (als Beispiel sei auf die Seite zu Johann Wolfgang von Goethe verwiesen: https://www.wikidata.org/wiki/Q5879).

Unser Framework umfasst derzeit vier Schritte, die im Folgenden beschrieben werden:

Vorverarbeitung und Konvertierung der Daten von RDF/XML zu JSON (rdf2json.py)

RDF/XML wird von den üblichen Softwaretools im Allgemeinen nicht als Datenstrom verarbeitet, sondern im Hauptspeicher abgelegt und dann weiterverarbeitet.

Aufgrund der Größe der Daten scheidet diese Möglichkeit aus. Da jedoch alle wesentlichen Daten zu einem Medium typischerweise innerhalb eines XML-Tags "rdf:Description" abgelegt sind, können wir die Daten auch mit Hilfe eines SAX-Parsers als XML verarbeiten. Wir extrahieren die für die Analyse wesentlichen Metadaten (z. B. dcterms:contributor, dcterms:language, dc:title, dcterms:extent, rdau:P60493) und speichern diese als JSON ab. JSON ist im Allgemeinen platzsparender als RDF/XML und kann leicht in Elasticsearch eingeladen werden, was ein geplanter nächster Schritt ist.

Extraktion von Daten zu Autoren aus Wikidata (WKD-Toolkit)

Unser Ziel ist die Anreicherung der Autorenangaben im DNB-Datensatz mit Informationen aus Wikidata, beispielsweise Geburtsdatum- und #ort, Beruf und Verweis auf einen etwa vorhandenen Artikel in Wikipedia. Da die Python-Softwarebibliothek zur Verarbeitung von Wikidata-Datensätzen veraltet ist, greifen wir auf das Java-basierte Wikidata Toolkit zurück. Nach Herunterladen des aktuell (14.08.2017) 16 GB großen komprimierten Wikidata-Datensatzes extrahieren wir in zwei Durchgängen zunächst alle Elemente mit einer GND-Kennung einschließlich ausgewählter Merkmale und ergänzen im zweiten Durchlauf die Werte der Merkmale. Das Ergebnis speichern wir im JSON-Format.

Normalisierung und Anreicherung der Daten (json2json.py)

Unser Python-Skript implementiert eine Pipeline, die alle in den vorherigen Schritten extrahierten Daten einliest und mit Hilfe der GND-Kennung verknüpft, Metadatenangaben (wie z. B. Seitenanzahlen) extrahiert, vereinfacht und normalisiert, Datensätze mit fehlenden Angaben filtert und schließlich die gewünschten Datenfelder spaltenbasiert ausgibt. Die Vereinfachung umfasst vor Allem das Entfernen von Namespace-Präfixen (etwa http://id.loc.gov/vocabulary/iso639-2/ bei der Angabe der Sprache); Seitenanzahlen werden mit Hilfe eines regulären Ausdrucks extrahiert, der die häufigsten Fälle abdeckt; Jahreszahlen ebenso; Verlagsnamen können mit Hilfe einer Normtabelle normiert werden (dies ist nötig, da die Schreibung dieser Namen innerhalb des Katalogs nicht standardisiert ist).

Analyse der Daten (Shell-Skripte und -Tools wie awk, sort, datamash, gnuplot, ...)

Die entstandenen Dateien im TSV-Format können mit den üblichen Unix-Kommandozeilen-Werkzeugen wie awk, sort, uniq etc. leicht verarbeitet und analysiert werden; Visualisierungen wurden mit gnuplot erzeugt. Alle Schritte sind im GitHub-Repository dokumentiert.

Zeitliche Entwicklung über 105 Jahre DNB

Abbildung 1 zeigt die zeitliche Verteilung einiger Subdatensätze des Katalogs. Von den etwa 14,1 Mio. Objekten im originalen DNB-Datensatz weisen etwa 8,3 Mio. extrahierbare Seitenanzahlen auf (59 %). Beschränken wir diese Anzahl auf ›Romane‹ (über das Datenfeld "rdau:P60493"), bleiben 353 498 übrig, von denen wiederum 316 518 Umfangsangaben aufweisen und 180 219 einen Verfasser, der mindestens einen Wikipedia-Eintrag (in egal welcher Sprache) besitzt. Dieses Datenset ist die Grundlage für die unten folgenden Anwendungsszenarien.

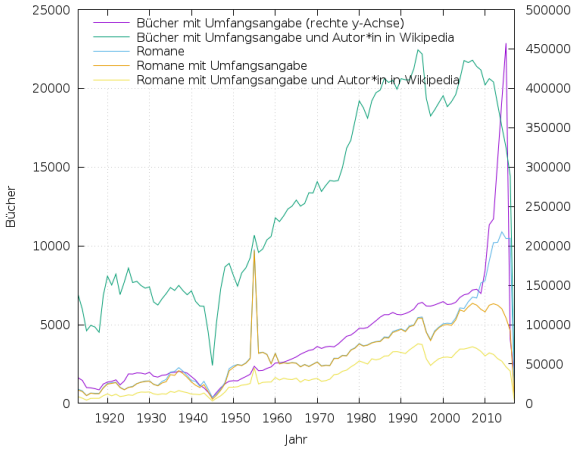

Abbildung 1: Fünf verschieden qualifizierte Subdatensätze des DNB-Katalogs in zeitlicher Verteilung.

Repräsentativität

Als möglicher Plausibilitäts- bzw. Repräsentativitätstest kann das Auszählen derjenigen Romanciers dienen, die mit den meisten Romanen im Katalog vertreten sind. Da der DNB-Katalog Vollständigkeit anstrebt, kann ein entsprechendes Ranking etwas über vergangene Realitäten auf dem deutschsprachigen Buchmarkt aussagen (Tab. 1), und tatsächlich stehen die Verfasser*innen von Romanbestsellern im Unterhaltungsbereich ganz oben (die Anzahl der Bücher umfasst von der DNB mitgesammelte Neuauflagen, Konsalik hat also nicht über 2 000 Romane geschrieben).

Autor*in	Romane
Heinz G. Konsalik	2232
Marie Louise Fischer	1264
Gert Fritz Unger	1013
Georges Simenon	783
Utta Danella	778
Edgar Wallace	654
Hedwig Courths-Mahler	647
Eleanor Hibbert	635
Pearl S. Buck	596
Alistair MacLean	582
Stephen King	577
Georgette Heyer	576
Agatha Christie	574
Theodor Fontane	565
Hans Ernst	563
Lion Feuchtwanger	501
Erich Maria Remarque	419
Hans Hellmut Kirst	411
Johannes Mario Simmel	403
Hans Fallada	396
Heinrich Mann	394
Fjodor Dostojewski	390
Barbara Cartland	390
Nora Roberts	381
Graham Greene	375
A. J. Cronin	370
Vicki Baum	366
Thomas Mann	359
Robert Ludlum	358
Gerd Hafner	357
Dean Koontz	354
Heinrich Böll	340
Alexandra Cordes	325
John le Carré	322
Marion Zimmer Bradley	321
Jason Dark	317
Willi Heinrich	313
Ludwig Ganghofer	311
Jack London	309
Joseph Roth	307
Danielle Steel	299
Johanna Lindsey	288
Erle Stanley Gardner	287
Siegfried Lenz	279
Jules Verne	277
Rosamunde Pilcher	274
Franz Kafka	271
Ernest Hemingway	271
Taylor Caldwell	269
Dorothy L. Sayers	269

Tabelle 1: Romanautor*innen geordnet nach Anzahl der Werke (inkl. Nachauflagen) im DNB-Katalog.

Anwendungsfall 1: Buchtitel

Die Verfügbarkeit großer digitalisierter Kataloge ermöglicht Large-Scale-Analysen bibliografischer Metadaten, etwa die Entwicklung von Romantiteln. Ein Vorläufer auf diesem Gebiet, Werner Bergengruens immer noch zu empfehlende Bibliothekarsfantasie »Titulus« von 1960, musste sich noch auf eine manuelle Sammlung des Autors stützen. Mittlerweile gibt es mit Franco Morettis Studie »Style Inc.« (2009) ein prominentes datengestütztes Beispiel (wobei sich Moretti bei seiner Analyse von um die 7 000 Romantiteln auf Fachbibliografien stützte, nicht auf Katalogdaten).

Um einen ersten Einblick in das Vokabular von Romantiteln zu bekommen, seien in Tabelle 2 die am häufigsten vorkommenden Substantive aufgelistet.

Substantiv	Frequenz
Liebe	3117
Mann	1906
Frau	1686
Tod	1537
Nacht	1505
Leben	1496
Welt	1188
Haus	1158
Zeit	1037
Schatten	1029

Tabelle 2: Häufigste Substantive in Romantiteln im gesamten DNB-Katalog.

Überzeitliche Konzepte – Liebe, Tod usw. – dominieren das Feld. Und nebenbei bemerkt: Ein wenig erinnert diese Liste an Jan Böhmermanns satirischen Song »Menschen, Leben, Tanzen, Welt«, mit dem auf die Beliebig- und Austauschbarkeit kontemporärer deutschsprachiger Liedproduktion angespielt wird (vgl. Pandzko/Böhmermann 2017), ein Befund, der sich analog auch auf Romantitel projizieren ließe.

Diese Anfragetechnik kann – wie beim Google Ngram Viewer – auf n-Gramme ausgedehnt werden, die Top-10 der häufigsten Trigramme findet sich in Tabelle 3.

Trigramm	Frequenz
Das Geheimnis der	238
Das Haus der	224
Der Mann der	189
Das Geheimnis des	175
Die Tochter des	160
Im Schatten des	128
Der Mann im	128
Das Lied der	125
Die Frau des	124
Die Reise nach	108

Tabelle 3: Häufigste Trigramme in Romantiteln im DNB-Katalog.

Ebenfalls analog zum Ngram Viewer lässt sich die zeitliche Entwicklung von n-Gramm-Frequenzen darstellen. Die unterschiedlichen Darstellungen in absoluten (Abb. 2) und relativen Zahlen (Abb. 3) kann etwa zeigen, dass sich zwischen Mitte der 1970er-Jahre und Mitte der 1990er-Jahre die Zahl an Romanen mit »Liebes«-Titeln zwar nahezu verdoppelt, dass sich diese Titel aber in relativen Zahlen nicht großartig vermehren.

Für genauere Analysen auf Grundlage dieser Extraktions- und Visualisierungsmethoden stellt das von uns vorgestellte Framework eine ideale Basis dar.

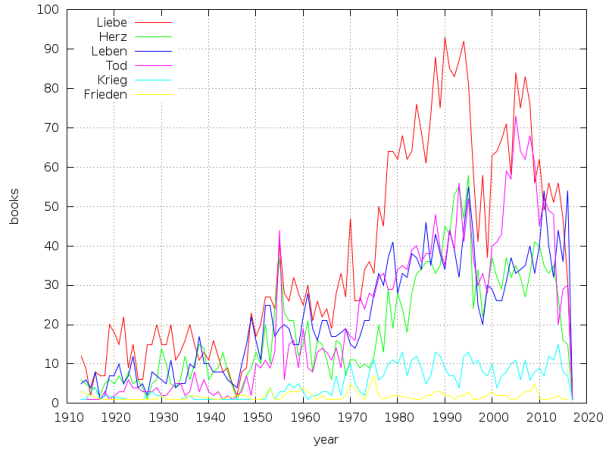

Abbildung 2: Vorkommen ausgewählter Wörter in Romantiteln im zeitlichen Verlauf (absolut).

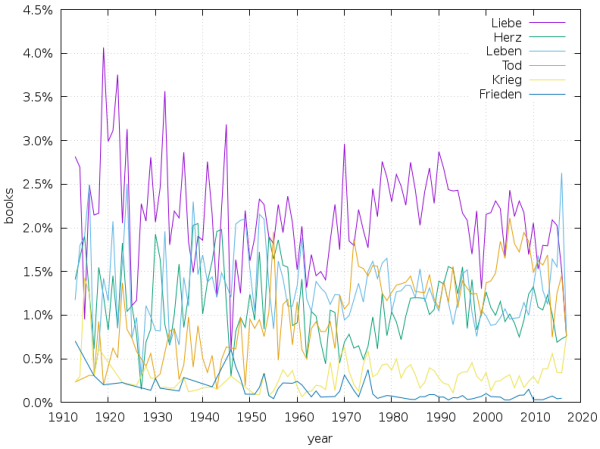

Abbildung 3: Vorkommen ausgewählter Wörter in Romantiteln im zeitlichen Verlauf (relativ).

Anwendungsfall 2: Textumfang

Unser zweites Anwendungsszenario betrifft die Erforschung des literarischen Textumfangs. Abbildung 4 zeigt die durchschnittliche Seitenanzahl von Romanen im Katalog der DNB.

Als Zuarbeit zu einer Theorie des literarischen Textumfangs haben wir mit dem von uns hier vorgestellten Framework in einer umfangreicheren Studie untersucht, wie sich der Umfang von Romanen etwa auf die Kanonbildung auswirkt (längere Romane, speziell solche von mehr als 1 000 Seiten Umfang, haben es leichter, in Kanonlisten zu landen). Außerdem ist es uns gelungen zu zeigen, wie umfangreiche Romane die DNA von Verlagen bestimmen können (vgl. Fischer/Jäschke 2018).

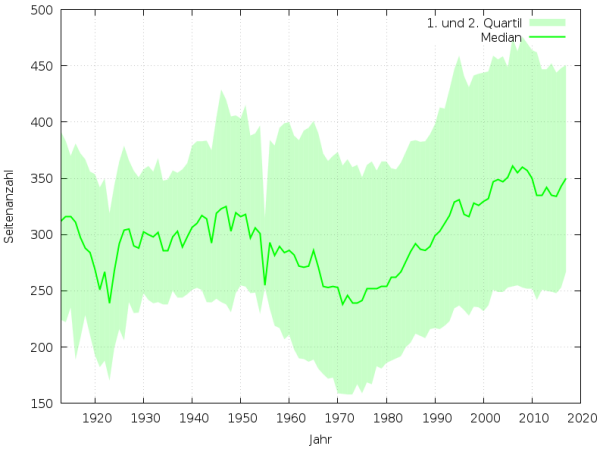

Abbildung 4: Entwicklung der mittleren Seitenanzahl pro Jahr seit 1913.

Fazit

Katalogdaten als Untersuchungsobjekt der quantifizierenden Literaturwissenschaften sind keine sich selbst erklärende Quelle, sondern ein über Jahrhunderte gewachsenes, überaus komplexes System. Die bibliothekarische Betreuung dieser Daten zielt nicht per se auf literaturwissenschaftliche Anwendungsfälle. Die Verschlagwortung kann lückenbehaftet sein, bestimmte Angaben wie etwa zum Textumfang können Fehler aufweisen. Die literaturwissenschaftliche Beschäftigung mit Katalogdaten setzt deren Explorier- und Kontrollierbarkeit voraus, wozu das hier vorgestellte Framework einen ersten Beitrag leisten soll. Zwei konkrete Anwendungsfälle sollten als Praxisbeispiele und ausdrücklich als Anreiz für weitere Szenarien dienen.

Bibliographie

Das **Arbeitsrepositorium** ist unter < https://github.com/weltliteratur/dnb > zu finden.

Bergengruen, Werner (1960): Titulus. Das ist: Miszellen, Kollektaneen u. fragmentar., mit gelegentl. Irrtümern durchsetzte Gedanken zur Naturgeschichte d. dt. Buchtitels oder unbetitelter Lebensroman e. Bibliotheksbeamten. Zürich: Verlag der Arche.

DNB (2017): »Wir über uns«, Stand 16.03.2017. URL: < http://www.dnb.de/DE/Wir/wir_node.html >.

Fischer, Frank; Jäschke, Robert (2018): Ein Quantum Literatur. Empirische Daten zu einer Theorie des literarischen Textumfangs. DFG-Symposium »Digitale Literaturwissenschaft«. Villa Vigoni, 9.–13. Oktober 2017. (Entsprechender Sammelband erscheint demnächst.)

Häntzschel, Günter; Hummel, Adrian; Zedler, Jörg (2009): Deutschsprachige Buchkultur der 1950er Jahre. Fiktionale Literatur in Quellen, Analysen und Interpretationen. Wiesbaden: Harrassowitz 2009. URL: < https://books.google.com/books?id=t88xc3CzK60C >.

Moretti, Franco (2009): Style Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850). In: Critical Inquiry, Vol. 36, No. 1 (Autumn 2009), S. 134–158.

Pandzko, Jim ; Böhmermann, Jan (2017): Menschen Leben Tanzen Welt [Musikvideo]. In: Neo Magazin Royale, 05.04.2017. URL: < https://youtu.be/h8MVXC_hqNY >.

Schmidt, Ben (2017): A brief visual history of MARC cataloging at the Library of Congress. In: Sapping Attention [Blog], 16.05.2017. URL: < http://sappingattention.blogspot.de/2017/05/a-brief-visual-history-of-marc.html >.