Linked Data an der UB Dortmund

Szenarien, Erfahrungen und ein Blick in die Werkstatt mit BiblioGraph

Hans-Georg Becker

Abstract

Die Universitätsbibliothek Dortmund stellt ihre bibliographischen Daten seit 2011 als Open Data zur Verfügung. Resultierend aus dem von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekt „ArcheoInf", an dem die UB Dortmund beteiligt war, gab es einige Ansätze, die bibliographischen Daten des Bibliotheksbestands auch als Linked Data zu modellieren und bereitzustellen. Das Ziel bei den Ansätzen war dabei immer, die Transformation vom klassischen Katalogisat hin zu einer semantischen Darstellung der Literatur mit Verlinkungen zu anderen Wissensdomänen voranzutreiben.
Der vorliegende Beitrag stellt die Ziele und Szenarien der UB Dortmund im Rahmen der Linked Open Data-Strategie vor. Dabei werden sowohl Erfolge als auch Probleme und Herausforderungen dargestellt. Neben einer Gegenüberstellung der Linked Data-Quellen, in denen die Daten der UB Dortmund bereits enthalten sind, gibt der Beitrag einen Einblick in die aktuellen Entwicklungen mit dem Produkt BiblioGraph von EBSCO.

Dortmund University Library has made its bibliographic data available as Open Data since 2011. As a result of the „ArcheoInf" project funded by the German Research Foundation (DFG), in which Dortmund University Library was involved, there were several approaches to modeling and providing the bibliographic data of the library collection as Linked Data.
The goal of these approaches was always to advance the transformation from the classic catalog to a semantic representation of the literature with links to other knowledge domains.
This article presents the goals and scenarios of Dortmund University Library in the context of its Linked Open Data strategy. Thereby, successes as well as problems and challenges are presented. In addition to a comparison of Linked Data sources that already contain the data of Dortmund University Library, the article provides an insight into the current developments with the product BiblioGraph from EBSCO.

Die Universitätsbibliothek Dortmund stellt ihre bibliographischen Daten seit 2011 als Open Data zur Verfügung.[1] Bereits seit 2010, als Folge des von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts „ArcheoInf", an dem die UB Dortmund beteiligt war[2], gab es einige Ansätze, die bibliographischen Daten auch als Linked Data zu modellieren und bereitzustellen.
Das Ziel bei den Ansätzen war dabei immer, die Transformation vom klassischen Katalogisat hin zu einer semantischen Darstellung der Literatur mit Verlinkungen zu anderen Wissensdomänen voranzutreiben. Als Grundlage dienen die Konzepte der Functional Requirements for Bibliographic Records (FRBR) sowie des CIDOC Conceptual Reference Model (CRM)[3] als Referenzontologie. Der Vorteil bei der „FRBRisierten" Sichtweise auf die bibliographischen Daten liegt vor allem in der Bündelung der teilweise zahlreichen Manifestationen und Expressionen hin zu einer Darstellung des Werks mit allen seinen untergeordneten Erscheinungsformen. So kann in Bibliothekskatalogen und Discovery Systemen das „Rauschen" in den Trefferlisten deutlich reduziert und die strukturelle Aussagekraft des Treffers erhöht werden.
Zusätzlich zur Modellierung der bibliographischen Daten als Linked Data, verfolgt die UB Dortmund das Ziel, mit der Öffnung der Daten auch Teil einer im Web verteilten Datenbank zu sein, die es ermöglicht, offene Daten unterschiedlicher Domänen miteinander zu verknüpfen, aber auch im Sinne einer Datenbank dynamisch abzufragen.
Der vorliegende Beitrag stellt die Ziele und Szenarien der UB Dortmund im Rahmen der Linked Open Data-Strategie vor. Dabei werden sowohl Erfolge als auch Probleme und Herausforderungen dargestellt. Ferner gibt der Beitrag einen Einblick in die aktuellen Entwicklungen mit dem Produkt BiblioGraph von EBSCO.

Ziele bei der Bereitstellung von Linked Open Data

Bei der Bereitstellung der bibliographischen Daten als Linked Open Data verfolgt die UB Dortmund fünf Ziele, die im Folgenden kurz erläutert werden.

Das erste Ziel ist, die bibliographischen Daten so zu modellieren und aufzubereiten, dass eine saubere hierarchische Repräsentation der Werke mit allen Expressionen und Manifestationen entsteht („FRBR-WEM-Struktur").
Das zweite Ziel ist die Transformation der bibliographischen Daten der UB Dortmund in das entwickelte Application Profile. Um dabei unnötige dublette Werke und Expressionen zu verhindern, ist es notwendig, die biblio-

1 https://data.ub.tu-dortmund.de/ [16. April 2023]

2 https://data.ub.tu-dortmund.de/archeoinf/ [16. April 2023]

3 Beim CIDOC Conceptual Reference Model handelt sich um eine Norm (ISO 21127:2006) für den kontrollierten Austausch von Informationen im Bereich des kulturellen Erbes. Mit dem CIDOC CRM wird das Ziel verfolgt, die vielfältigen Informationen im Bereich des kulturellen Erbes gemeinsam zu erfassen und einen allgemeinen Rahmen ihrer formalen Semantik zur Verfügung zu stellen, damit jede Information dieses Bereichs den Begriffen des CIDOC CRM zugeordnet werden kann.

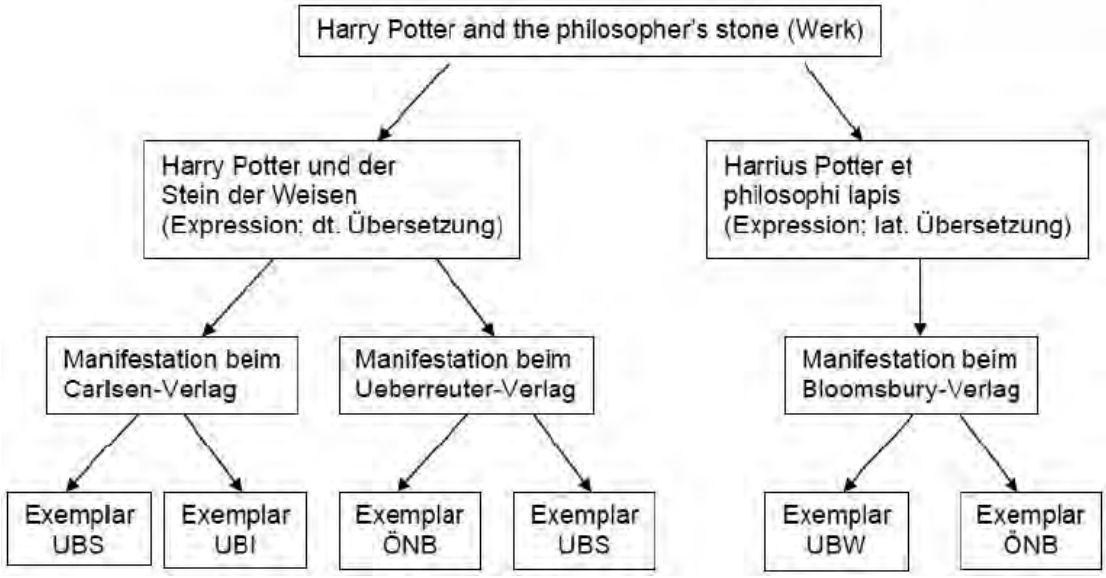

Abbildung 1: Graph eines Werkes mit all seinen Ausprägungen

graphischen Metadaten zu analysieren und mittels in den Daten enthaltener Fakten aber auch durch Heuristiken zu bündeln, anzureichern und zu verlinken.[4] Diese Aufgabe ist nicht trivial. Bisherige Ansätze erzeugen immer auch falsch gebündelte und falsch verlinkte Strukturen.

In Abbildung 1 ist dargestellt, wie ein Graph in FRBR-WEM-Struktur aussehen könnte. Es handelt sich hierbei um eine Serie mit zwei Bänden, wobei der zweite Band sowohl als gedruckte als auch als digitale Ausgabe erschienen ist. Durch die in den Quelldaten erstellte Verknüpfung entstehen für den zweiten Band zwar vier Manifestationen, jedoch werden diese passend mit „same as"-Relationen verknüpft. Das Bild und die Beschreibung müssen noch besser werden, da es ein BIBFRAME-Bild ist und nicht FRBR.

Das dritte Ziel ist die Veröffentlichung der Daten als Linked Open Data als Teil einer im Web verteilten semantischen Datenbank. Dazu reicht es nicht aus, die Daten als Download zur Verfügung zu stellen. Vielmehr müssen die Daten in geeigneten offenen und standardisierten Datenbanken mit der Möglichkeit zum dynamischen Abfragen der Daten bereitgestellt werden. Hierzu dienen sogenannte Triple Stores, die auch SPARQL-Schnittstellen zur Verfügung stellen.

Als viertes Ziel möchte die UB Dortmund erreichen, dass die bibliographischen Daten Teil der Knowledge Graphs der großen Suchmaschinen werden. Dazu ist es notwendig, dass im Application Profile auch Klassen und Eigenschaften der Ontologie schema.org enthalten sind.[5]

Last but not least sollen die entstandenen Daten auch in Anwendungen wie dem Katalog bzw. Discovery Service (weniger „Rauschen" und mehr Struktur) oder der virtuellen Systematik (Anwendung von inhaltlichen Erschließungselementen auf nicht erschlossene Entitäten) nachgenutzt werden.

Erfolge und Herausforderungen

Bisher konnte das erste Ziel, die Erstellung eines für die UB Dortmund passenden Application Profile, umgesetzt werden.[6] [7] Das entwickelte Datenmodell basiert auf dem CIDOC CRM unter Verwendung der Erweiterung FRBRoo als Referenzontologie. Die Beschreibung der Entitäten erfolgt dabei unter Nachnutzung gängiger bibliographischer Ontologien bzw. Application Profiles, wie z.B. dem Application Profile von lobid-resources[8]. Bei der Transformation der bibliographischen Daten in das Application Profile zeigen sich durchaus Probleme bei der Zuordnung der einzelnen Felder. Das gilt insbesondere für die Unterscheidung von Eigenschaften für Werke und Expressionen. Da das Application Profile prozess- und ereignisorientiert ist, stellt sich auch die Zuordnung von Daten zu den Ereignissen als Herausforderung dar. Hierbei ist insbesondere auf die Vermeidung von „Blank Nodes" zu achten, da diese insbesondere im Szenario der dynamischen Abfrage von Daten mittels SPARQL zu Problemen führen

4 Das auf FRBR basierende Katalogisierungsregelwerk „Resource Description and Access" (RDA) stellt zwar grundsätzlich die Mittel zur Abbildung der WEM-Struktur vor, jedoch ist es unrealistisch, dass sämtliche Katalogdatensätze retrospektiv auf RDA angepasst werden. Ferner werden aufgrund der hohen Schlagzahl an neuen Publikationen vermehrt bibliographische, nicht RDA-konforme Daten von Verlagen direkt in die Kataloge der Bibliotheken eingespielt. Somit müssen Algorithmen und Heuristiken diese Aufgabe übernehmen. Ggf. kann eine kooperative manuelle Qualitätskontrolle ergänzt werden.

5 Wikipedia: „Die von Schema.org vorgeschlagenen strukturierten Daten können dazu verwendet werden, Suchmaschinen das Verständnis für den Kontext und die Bedeutung des Inhalts einer Webseite zu erleichtern, was die Darstellung der Seite in den Suchmaschinenergebnissen verbessern kann." https://de.wikipedia.org/wiki/Schema.org [16. April 2023]

6 Becker, Hans-Georg: FRBR, Serials und CIDOC CRM – Modellierung von fortlaufenden Sammelwerken unter Verwendung von FRBRoo, in: Patrick Danowski / Adrian Pohl (Hrsg.): (Open) Linked Data in Bibliotheken, Berlin 2013, S. 64-96. https://doi.org/10.1515/9783110278736.64 [16. April 2023]

7 Becker, Hans-Georg: Bestandsnachweise mit einem CIDOC CRM-Application Profile. https://the-lodlam-mercury.de/2014/08/10/bestandsnachweise-mit-einem-cidoc-crm-application-profile/ [16. April 2023]

8 https://lobid.org/resources/dataset [16. April 2023], Application Profile: https://lobid.org/resources/api#jsonld [16. April 2023]

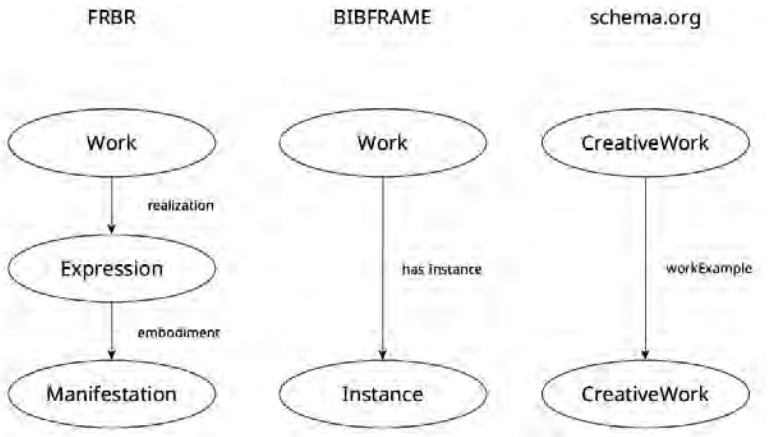

Abbildung 2: FRBR vs. BIBFRAME vs. Schema.org

können. Weitere Probleme ergeben sich bei der Bündelung von Werken und Expressionen zur Vermeidung von Dubletten. Da in den Quelldaten der klassischen Bibliothekssysteme und in den tatsächlich angewendeten Regelwerken der Formalerschließung kaum Eigenschaften erfasst werden, die die WEM-Struktur abbilden, sind für die Bündelung Heuristiken notwendig. Eine vermeintlich simple Heuristik ist die Bündelung über identische ISBNs. Allerdings sind ISBNs nicht so eindeutig, wie man meinen könnte. Beispielsweise werden von Verlagen ISBNs alter Werke nachgenutzt und neu vergeben. Aber auch die Vergabe identischer ISBNs für alle Teile einer Werkausgabe ist möglich. Somit kommt es bei der alleinigen Nutzung der ISBNs zu falschen Bündeln. Ferner können ISBNs nur als Kriterium für Werke dienen, die seit der Einführung der ISBN Anfang der 1970er Jahre erschienen sind.

Weitere Erfolge bei der Umsetzung der Ziele sind auf der rein technischen Ebene zu verzeichnen. So wurden etwa potentielle Triple Stores ausprobiert sowie Prozesse für den Import-, die Aktualisierung und die Löschung von Daten entwickelt. Allerdings stehen, durch das Fehlen geeigneter Heuristiken und Algorithmen, die für die Veröffentlichung geeigneten Daten noch nicht zur Verfügung.

Die FRBR-WEM-Sicht wird in der Community als problematisch und zu komplex wahrgenommen.[9] Insbesondere die Interpretation und die Ermittlung der Werke und Expressionen ist dabei Gegenstand der Diskussionen.

Einen etwas in der Komplexität gegenüber FRBR reduzierteren Ansatz verfolgt BIBFRAME. Bei der Entwicklung von BIBFRAME wurde versucht, die positiven Eigenschaften und die Probleme der FRBR-WEM-Struktur unter einen Hut zu bekommen. Der Ansatz bei BIBFRAME ist es, Werke und Expressionen aus FRBR in einer Klasse „Work" zusammenzufassen, um so die Probleme bei der Unterscheidung, was ein Werk und was eine Expression ist, zu umgehen.

Im schema.org-Vokabular wird noch weiter reduziert, sodass hier nur noch das Werk übrigbleibt („ComplexWork"). Somit ist man bei schema.org eigentlich wieder da, wo man mit den klassischen Bibliothekskatalogen schon war. Jedoch sieht schema.org vor, dass Werke mit der Relation „workExample" in ähnlicher Form in Verbindung stehen, wie Werke und Instanzen in BIBFRAME.

Abbildung 2 stellt die drei Strukturvarianten noch einmal grafisch gegenüber.

Da auch bei den bisherigen Ergebnissen der UB Dortmund ähnliche Beobachtungen bei der Zuordnung von Eigenschaften zu Werken und Expressionen gemacht wurden, kann die BIBFRAME-WI-Struktur ein praktikabler Kompromiss zur FRBR-WEM-Struktur sein. Die sich daraus ergebende Inkompatibilität zum CIDOC CRM/FRBRoo-Ansatz beim Application Profile, lässt sich jedoch technisch umgehen und gewährleistet so die Interoperabilität.[10] [11]

Bei all den bisherigen Ansätzen und Bemühungen hat sich auch gezeigt, dass die Entwicklung und Bereitstellung von bibliographischen Linked Open Data sowohl in Bezug auf Personal- als auch IT-Ressourcen aufwändig ist. Letztlich ist auch daher bei der Umsetzung der oben genannten Ziele in der UB Dortmund noch Luft nach oben.

Allerdings sind die bibliographischen Daten der UB Dortmund durch das Hochschulbibliothekszentrum (hbz) bzw. durch dessen Service „lobid-resources" als Linked Open Data veröffentlicht. Auch über andere Services stehen die Daten der UB Dortmund zur Verfügung. Neben lobid-resources, werden im Folgenden drei weitere Services beschrieben und bzgl. der Ziele eingeordnet.

lobid-ressources

Bei lobid-resources handelt es sich um ein Abbild des hbz-Verbundkatalogs mittels eines hbz-eigenen Application Profile, welches sich seit den Anfängen von lobid-resources deutlich weiterentwickelt hat. Nachdem sich das Application Profile anfangs sehr pragmatisch entwickelt hat, wurde es 2017 mit der Version 2.0 systematisch mit einer Definition der verwendeten Vokabulare und Ontologien überarbeitet. Die Basis legen die Entitäten bzw. Klassen der Bibliographic Ontology (BIBO). Zur Beschreibung der Entitäten kommen u.a. auch schema.org, BIBFRAME und RDA

9 Siehe beispielsweise Kapitel 9 in „FRBR Before and After" von Karen Coyle, 2016, https://kcoyle.net/beforeAndAfter/ [23. Mai 2023]

10 Eric Miller: Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services. Library of Congress, 21. November 2012, abgerufen am 28. Mai 2014. https://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf

11 Sofia Zapounidou, Michalis Sfakakis, Christos Papatheodorou: Metadata and Semantics Research. Hrsg.: Emmanouel Garoufallou, Jane Greenberg. Springer, Thessaloniki 2013, ISBN 978-3-319-03436-2, Highlights of Library Data Models in the Era of Linked Open Data, S. 396-407. Postprint: http://eprints.rclis.org/32103/1/mtsr2013_postprint.pdf

zur Anwendung.[12] Jedoch entsteht so keine echte an FRBR angelehnte Struktur. Allerdings werden die sogenannten Einheitssachtitel und optional die GND-ID eines Werks angegeben, wenn die Ressource eine Instanz enthält. Diese Information lässt sich bei den Heuristiken zur Bündelung heranziehen. Die Daten haben noch weitere Einschränkungen bei der Verwendung. Zum einen können die Daten nur als JSON-LD und JsonLines[13] heruntergeladen oder einzelne Ressourcen mittels Content Negotiation[14] aufgerufen werden, was einige Nacharbeiten zur Folge hat. Zum anderen enthalten die Daten aufgrund des Application Profiles einige Blank Nodes, die SPARQL-Abfragen erschweren.

Die UB Dortmund nutzt bereits seit einigen Jahren ihren Datenanteil aus lobid-resources für die Indexierung im Discovery Service und zur Live-Präsentation der Detailansicht von Suchtreffern in „Katalog plus". Hierzu wird die JSON-Version der Daten verwendet. Diese Daten werden auch im Rahmen der virtuellen Systematik in Katalog plus verwendet, um die Daten aus der folgend beschriebenen Quelle „CultureGraph" anzureichern.

CultureGraph

CultureGraph[15] ist eine Zusammenstellung aller bibliographischen Daten aus den deutschen Verbünden, der Deutschen Nationalbibliothek und des österreichischen Verbunds. Ziel von CultureGraph ist die Bildung von Werkclustern, sodass insbesondere eine Anreicherung und Übernahme mit Erschließungselementen, wie Sacherschließung und GND-Verknüpfungen, erreicht werden kann. Die UB Dortmund nutzt diese Daten als Grundlage für die virtuelle Systematik auf Basis der RVK im Discovery System, da im hbz-Verbund kaum RVK-Notationen vorhanden sind. Ferner werden GND-Verknüpfungen übernommen.[16]

Die Daten sind mittels MARC 21 beschrieben und beinhalten keine vollständigen bibliographischen Datensätze, sondern reduzieren sich auf die Daten zu den Clustern. Für eine vollständige Sicht auf die bibliographischen Daten werden also weitere Daten benötigt. Diese Tatsache und die Größe der Daten-Dumps von CultureGraph resultieren in einem hohen lokalen technischen Aufwand für die Nachnutzung der Daten. Unter anderem hat die UB Dortmund mittels der „Bundle Ontology"[17] die Daten in RDF umgewandelt, damit sie mittels SPARQL nutzbar werden. Ferner werden die Daten mit einer simplen RDF-Version der lobid-resources im Triple Store verknüpft. Die CultureGraph-Cluster können ebenfalls als Heuristik für die Bündelung der Werke herangezogen werden.

Bibliotheksmanagementsystem Alma

Seit 2019 steigt der hbz-Verbund auf Alma von ExLibris um. Die UB Dortmund gehörte 2021 zur ersten Welle der umsteigenden Bibliotheken. Über Alma lassen sich bibliographische Daten mit unterschiedlichen Verfahren und in einer Reihe von Formaten exportieren. Unter anderem besitzt Alma ein sogenanntes Publishing Profile für BIBFRAME-Daten. Derzeit ist der Export jedoch fehlerhaft und kann nicht in Gänze analysiert werden. Einzig kann festgehalten werden, dass beim Export zwar die BIBFRAME-WI-Struktur geliefert wird, jedoch eine Reduzierung von Werken mittels Cluster-Verfahren nicht vorgesehen ist. Ferner entstehen auch hier einige Blank Nodes, die eine Nachnutzung via SPARQL problematisch machen.

OCLC Worldcat

Worldcat ist der größte bibliographische Katalog der Welt, in den auch Daten der deutschen Verbünde und somit auch die Daten der UB Dortmund einfließen. Die Daten im Worldcat werden über Importverfahren aus den Bibliotheken und Bibliotheksverbünden gesammelt, wobei naturgemäß viele Dubletten entstehen. OCLC hat daher schon früh damit begonnen, Verfahren zu Bündelung von Datensätzen zu entwickeln und einzusetzen. Ferner wurde im Worldcat schon früh auf das Vokabular schema.org gesetzt und OCLC hat wesentlich dazu beigetragen, dass in schema.org auch Werk-Cluster abgebildet werden können. Der Worldcat hat auch viele Jahre Linked Data bei seinen Datensätzen angeboten, die auch mittels Format-Suffix an der URL abrufbar waren. Daten-Dumps waren nie vorgesehen. Die Lizenz war allerdings immer sehr eingeschränkt (CC-BY-NC-ND) und kein echtes Open Data. Seit einiger Zeit bietet OCLC im Worldcat die Linked Data-Repräsentation nicht mehr an, womit der relevante Datenanteil in einem weiteren Silo versackt ist.

Zwischenfazit

Selbst wenn die Daten der UB Dortmund in einem brauchbaren Application Profile inkl. Datenoptimierung im FRBR/BIBIFRAME-Sinne vorliegen würden, so gäbe es trotzdem noch keine im Web verteilte semantische Datenbank,

12 https://blog.lobid.org/2017/04/19/vocabulary-choices.html [3. Juli 2023]

13 JsonLines ist ein Format, bei dem pro Zeile ein JSON-Datensatz ausgegeben wird. Dieses Format ist in vielen Fällen ohne Nacharbeit nicht sofort nutzbar.

14 Seite „Content Negotiation". In: Wikipedia: Die freie Enzyklopädie. Bearbeitungsstand: 20. Juni 2023, 10:23 UTC. URL: https://de.wikipedia.org/w/index.php?title=Content_Negotiation&oldid=234771527 (Abgerufen: 3. Juli 2023, 09:24 UTC)

15 https://www.dnb.de/DE/Professionell/Standardisierung/AGV/_content/culturegraph_akk.html [3. Juli 2023]

16 Mein Artikel zur virtuellen Systematik

17 Die „Bundle Ontology"ist derzeit leider nicht offen im Internet veröffentlicht.

Abbildung 3: BiblioGraph-Uploader

also offen zugängliche Triple Stores, die mittels Federated SPARQL[18] abgefragt werden könnten. Somit müssen weiterhin alle notwendigen Daten lokal heruntergeladen, ggf. aufwändig aufbereitet und in einen lokalen Triple Store geladen werden. An der Ressourcenproblematik ändert sich damit also nichts.

Es stellt sich daher die Frage, ob ein externer Dienstleister beim Erreichen der Ziele helfen kann und was die Erwartungen an einen Service wären.

Benötigt würde zunächst eine Repräsentation der Daten der UB Dortmund mindestens als BIBFRAME und ohne eigene Anpassungen machen zu müssen. Ferner wäre eine intelligente Anreicherung der Daten zu Werk-Clustern und damit einhergehend eine Reduktion von Werken gewünscht. Die Integration der Daten in die „Knowledge Graphs" der Suchmaschinen und somit eine Repräsentation mittels schema.org-Vokabular ist ebenso gefordert, wie eine offene Lizenz, optimaler Weise „public domain". Die Bereitstellung der Daten als Dump sowie die Bereitstellung eines Triple Stores inkl. SPARQL-Endpoint runden die Anforderungen ab.

Aktuelle Entwicklungen mit BiblioGraph

Im Oktober 2022 bekam die UB Dortmund die Anfrage von EBSCO, ob Interesse als „Early Adopter" an BiblioGraph bestünde. In einer Videokonferenz wurden Details zum Service vorgestellt und die nächsten Schritte vereinbart.

Bei BiblioGraph handelt sich in erster Linie um einen Service, der aus klassischen bibliographischen Metadaten Linked Data im Bibframe-Modell erzeugt. Auch eine Anreicherung mit anderen offenen Daten ist dabei möglich. Die Daten können zum einen in Folio[19] verwendet werden, werden aber auch in einer dafür bereitgestellten Datenplattform von EBSCO gehostet. Ein weiterer Aspekt ist die „Google-Syndication". Hier wird das Ziel verfolgt, dass auf Wunsch der beteiligten Bibliothek die Daten in den Knowledge Graphs der großen Suchmaschinen bereitgestellt und so die Bibliotheken und deren Bestände in Google & Co. sichtbarer werden. Die reine Indexierung der Daten in den Suchmaschinen wird direkt über die BiblioGraph-Datenplattform realisiert. Die in BiblioGraph entstandenen Daten sind offen und mindestens unter CC-BY 4.0 und demnächst auf Wunsch auch als CC0 lizenziert. BiblioGraph bietet auch Services auf Basis der entstandenen Daten an, die in eigene Webseiten eingebunden werden können. Diese Services sind nicht Gegenstand des vorliegenden Beitrags, da diese das Erreichen der oben genannten Ziele nicht beeinflussen.

Um BiblioGraph nutzen zu können, wird zunächst ein möglichst aussagekräftiger Export der eigenen bibliographischen Daten benötigt. Derzeit wird von BiblioGraph nur MARC-XML unterstützt. Die Unterstützung weiterer Formate ist in Planung.

Das Bibliotheksmanagement Alma bietet die Möglichkeit, die Daten als MARC-XML zu exportieren und ggf. bereits beim Export zu filtern oder sogar anzupassen und anzureichern (z.B. mit Daten aus in Relation stehenden Datensätzen). Die Daten werden dann in BiblioGraph hochgeladen. Das kann sowohl manuell als auch mittels API geschehen. Abbildung 3 zeigt den Uploader für die Daten.

18 Federated SPARQL Queries sind SPARQL-Queries, die über eine oder mehrere SPARQL-Endpoint verteilt arbeiten. So ist es z.B. möglich, Daten aus dem eigenen Triple Store mit Daten aus Wikidata oder anderen öffentlichen Datenquellen zu verknüpfen. S.a. https://www.w3.org/TR/sparql11-federated-query/ [3. Juli 2023]

19 https://de.wikipedia.org/wiki/Folio_(Bibliothekssoftware)

www.narr.digital

Unsere eLibrary bietet Ihnen aktuell rund 1.900 eBooks in den Formaten ePDF und ePub sowie etwa 8.900 Artikel aus über 500 Zeitschriftenausgaben.

Sie ist an den OCLC WorldCatalogue angebunden.

Ihre Ansprechpartnerin für Campuslizenzen:
Claudia Marcks (marcks@narr.de)

Profitieren Sie von unseren Lizenzmodellen

\ Paketkäufe, Pick&Choose-Erwerbungen und Einzelabonnements individuelle Angebote auf Anfrage

\ EBS-Angebote für das gesamte Programm der Verlage Narr Francke Attempto, UVK und expert

narr\f ranck e\atte mpto

UVK

expert

vernarrt in eBooks.

Alle Vorteile auf einen Blick

\ Zugriff über IP-Adressen, VPN oder Shibboleth
\ unbeschränkte Nutzerzahl
\ Downloadzahlen (COUNTER, SUSHI)
\ Bestandsübersichten (Marc-XML, KBART)
\ keine Gebühren
\ keine Mindestbestellmenge
\ weiches DRM
\ Erwerbungsvorschläge
\ Administratorenzugang

Abbildung 4: BiblioGraph-Dashboard: Historie

Details / History

Copy CSV Excel Print

Date	Database	Pipeline	Library Link Domain	Library Locations	MARC Records	Bibframe Resources	Links
2022-12-08T10:38:30Z	v7	7.0.4	https://tudortmund.library.link/	1	1501794	6221627	25149672
2022-12-07T15:06:36Z	v7	7.0.4	https://tudortmund.library.link/	1	1501794	6221627	25149672
2022-10-06T06:12:31Z	v7	7.0.4	https://tudortmund.library.link/	1	1481766	6015697	24030770
2022-09-01T01:34:47Z	v7	7.0.4	https://tudortmund.library.link/	1	1481767	6015699	24030774
2022-08-30T03:07:22Z	v7	7.0.4	https://tudortmund.library.link/	1	1481766	6015697	24030770
Date	Database	Pipeline	Library Link Domain	Library Locations	MARC Records	Bibframe Resources	Links

Showing 1 to 5 of 5 entries

Library Resource Graph

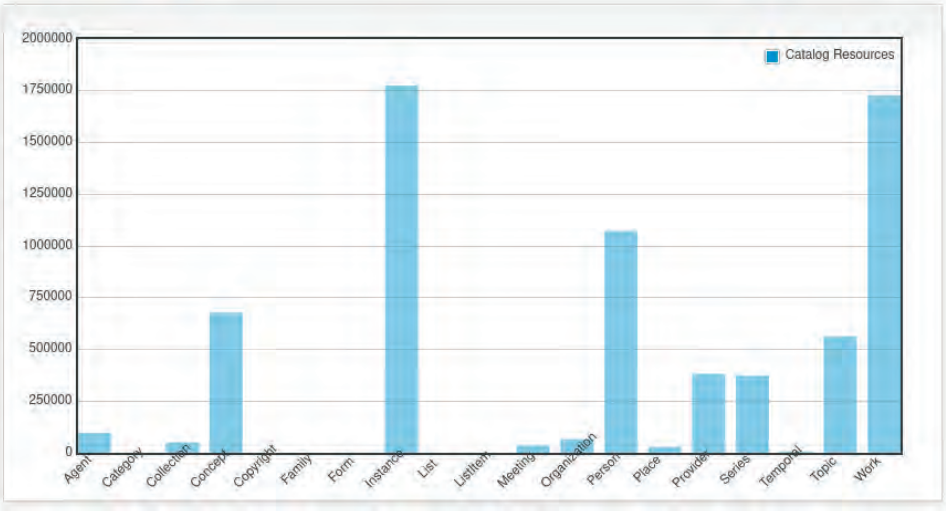

Abbildung 5: BiblioGraph-Dashboard: Knoten des Graphs

Nach dem Upload startet ein Datenprozess, auf dessen Abschluss zu warten ist. Dies kann einige Tage dauern. Nach Abschluss des Prozesses lässt sich das Ergebnis in einem Dashboard (s. Abbildung 4) betrachten.

Das Dashboard bietet auch Angaben zu den erzeugten Ressourcen im Graph (s. Abbildung 5). Bei der Analyse der entstandenen Ressourcen ist sofort ersichtlich, dass sich die Anzahl der „Works" nur geringfügig von der Anzahl der „Instances" unterscheidet. Somit ist zwar bei BiblioGraph eine an FRBR angelehnte Struktur entstanden, jedoch hat das Verfahren auch hier nicht zu einer wesentlichen Zusammenführung und Reduktion der Werke geführt.

Ein Blick in die Eigenschaften der Werke und Instanzen zeigt jedoch, dass sich durchaus neue Beziehungen gebildet haben, die zur Bündelung beitragen können. Für diese Detailanalyse wurde der UB Dortmund ein Abzug der Daten in RDF/XML zur Verfügung gestellt. Dieser Abzug wurde in eine lokale GraphDB von Ontotext eingespielt und mittels SPARQL und der Funktion „Visual Graph" erkundet. Die folgenden Beispiele zeigen sowohl positive Ergebnisse als auch Probleme.

Zunächst fällt positiv auf, dass keine Blank Nodes entstanden sind. Jeder Knoten im Graph hat somit eine Bezeichnung und kann direkt angesprochen werden. Ferner fällt auf, dass das Hauptkriterium für die Bildung von Werk-Clustern die ISBNs sind.

Im ersten Beispiel (s. Abbildung 6) wird schnell deutlich, warum auch die Daten im Original-MARC-Datensatz sauber sein sollten. Das Feld 020 für die ISBNs gibt es im Datensatz zu „FRBR, before and after" von Karen Coyle viermal mit ISBNs zu vier Manifestationen des Buches. Daraus macht der BiblioGraph-Algorithmus vier Instanzen und ein Werk in Bibframe. Das sieht auf den ersten Blick richtig aus. Leider erzeugt der Algorithmus auch zu jeder Instanz ein Exemplar, was falsch ist. Wäre die ISBNs zu den nicht lizenzierten Ausgaben im MARC-Feld 776 untergebracht, so wäre nur ein Exemplar entstanden. Da es durchaus vorkommen kann, dass im MARC-Feld mehr als eine ISBN zur erworbenen Manifestation gehören bzw. im Katalogisat mehr als eine Manifestation erfasst ist, ist das Verhalten von BiblioGraph grundsätzlich nicht falsch. Beispielsweise entsteht die Situation, wenn ein E-Book lizenziert wurde, welches in drei Formaten im Zugriff ist (etwa pdf, epub und html) und jedes davon eine eigene ISBN besitzt. Für diesen Fall muss auch zu jeder durch die ISBNs entstandene Instanz ein Exemplar erzeugt werden. In BiblioGraph sind die Exemplare allerdings nicht unterscheidbar und enthalten auch alle dieselbe Serviceinformation („borrowAction"). Hier werden also die durchaus im Upload mitgelieferten Daten sowohl für gedruckte als auch für elektronische Exemplare in BiblioGraph verwendet.

Am Beispiel „Semantic Web" (s. Abbildung 7) sieht man, dass BiblioGraph parallele Ausgaben sauber erkennt und korrekte „same as"-Beziehungen aufbaut. Die genaue Definition dieser „same as"-Beziehung bleibt bislang aber ungeklärt, da es sich um eine Relation aus der Biblio-

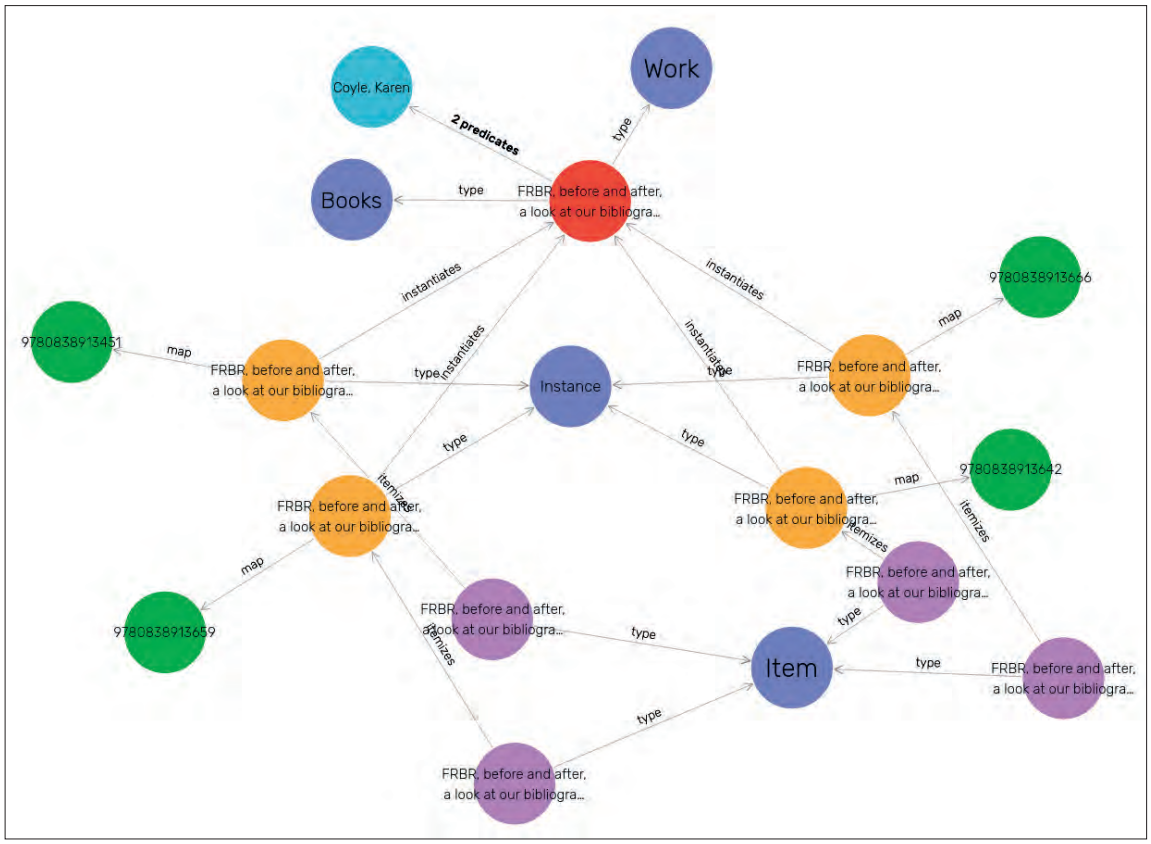

Abbildung 6

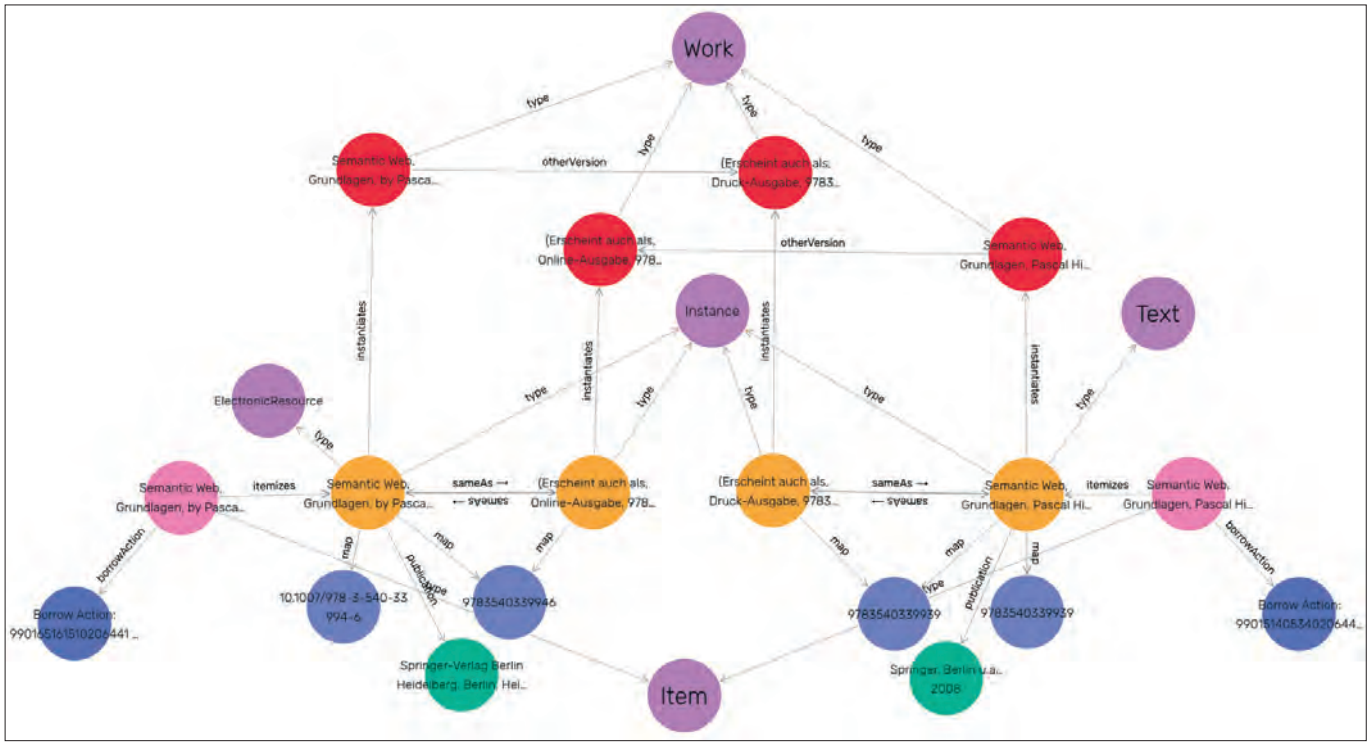

Abbildung 7: Semantic Web. Pascal Hitzler et.al.

Graph-eigenen Ontologie handelt und die Ontologie bisher noch nicht einsehbar ist. Auf der Werkebene sind mit „other Version"-Relationen ebenfalls Relationen zwischen den Ausgaben aufgebaut worden. Jedoch zeigt sich hier, dass für ein „echtes" Cluster noch zu viele Werke entstanden sind. Es gibt hier zu jeder Instanz ein Werk. Hier wäre eine Reduktion auf ein Werk zu begrüßen.

Wie bereits oben angemerkt sind ISBNs leider nicht so eindeutig, wie man sich das wünschen würde. Das hat sowohl Vor- als auch Nachteile. Im folgenden Beispiel zeigt sich ein Vorteil (s. Abbildung 8). Hier werden zwei Auflagen (Instanzen) eines Werkes miteinander verknüpft, da sie dieselbe ISBN erhalten haben. Aber auch hier werden zwei Werk-Entitäten erzeugt anstatt auf ein Werk zu reduzieren.

Aber so entstehen leider auch falsche Cluster, da der Algorithmus von BiblioGraph offenbar keine weiteren Heuristiken verwendet. In folgendem Fall werden zwei völlig

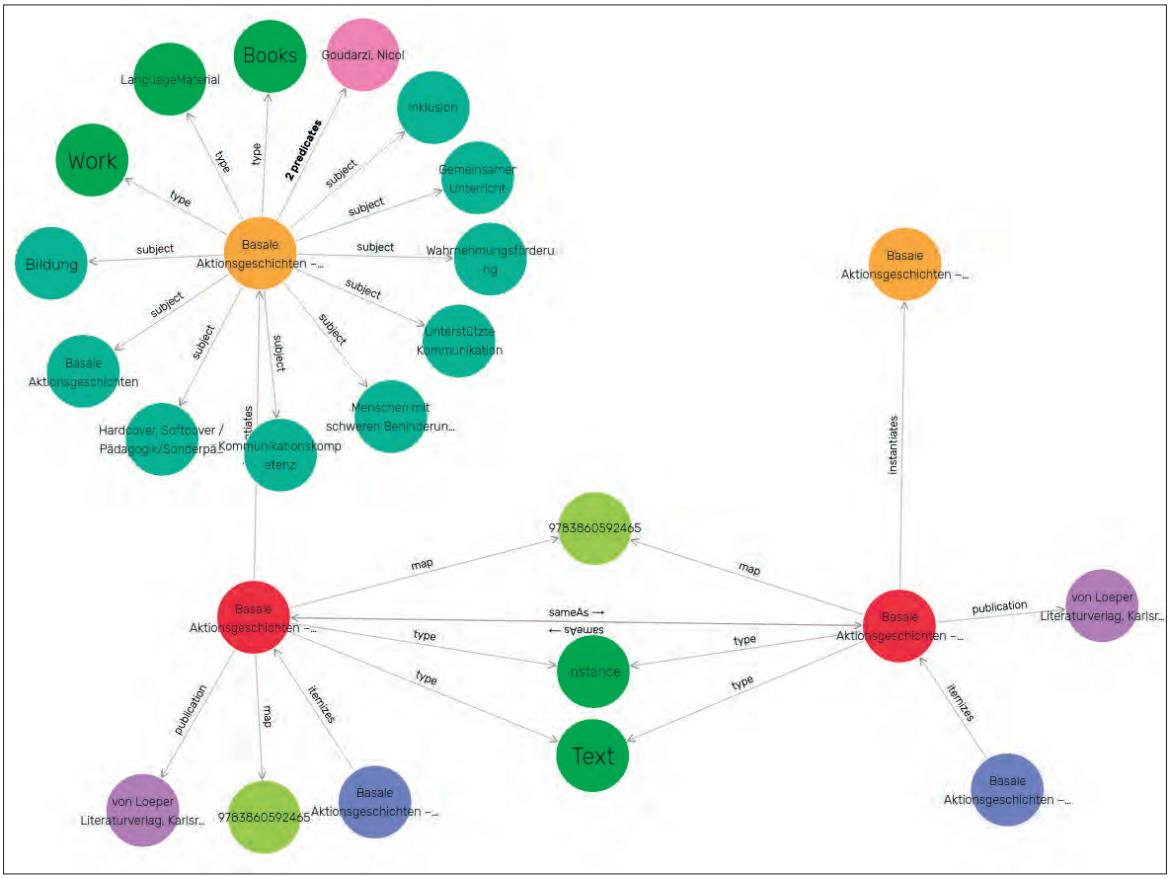

Abbildung 8: Basale Aktionsgeschichten – Vorteil ISBNs

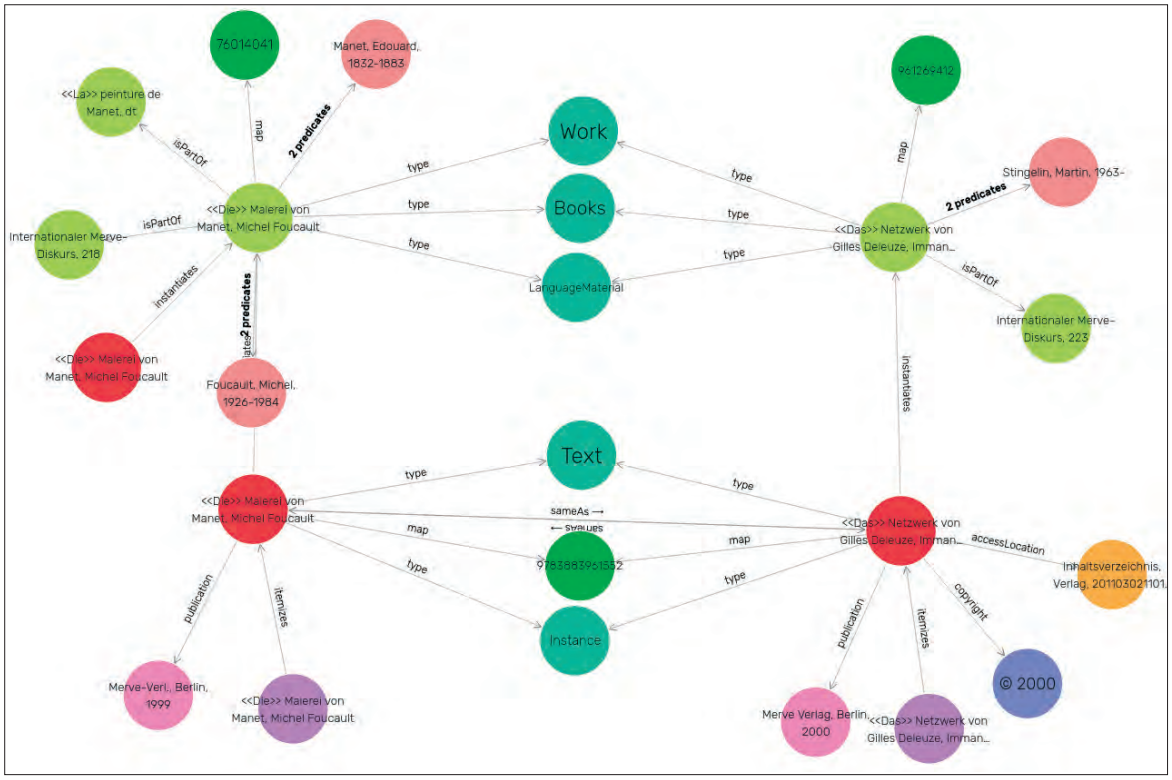

Abbildung 9: Nachteil ISBN

verschiedene Instanzen mit „same as“ in Relation gesetzt (s. Abbildung 9). Hierbei handelt es sich um zwei Werke einer Verlagsreihe, die, wie auch alle anderen Bände der Reihe, dieselbe ISBN erhalten haben.

Ein weiterer wichtiger Fall, bei dem eine gute Cluster-Bildung enorme Vorteile beim „Discovery“ hätte, sind mehrbändige Werke. Hier hat der Algorithmus von BiblioGraph schon einen sehr guten Ansatz gefunden, die Struktur des Gesamtwerks abzubilden (s. Abbildung 10). Der Algorithmus macht sich die Daten aus dem MARC-Feld 830 zunutze und findet alle Bände und verknüpft diese. Leider entstehen trotzdem schlechte Graphen, weil der Algorithmus das Unterfeld „v“ mit der Bandnummerierung nicht auswertet und so falsche „same as“-Relationen erzeugt.

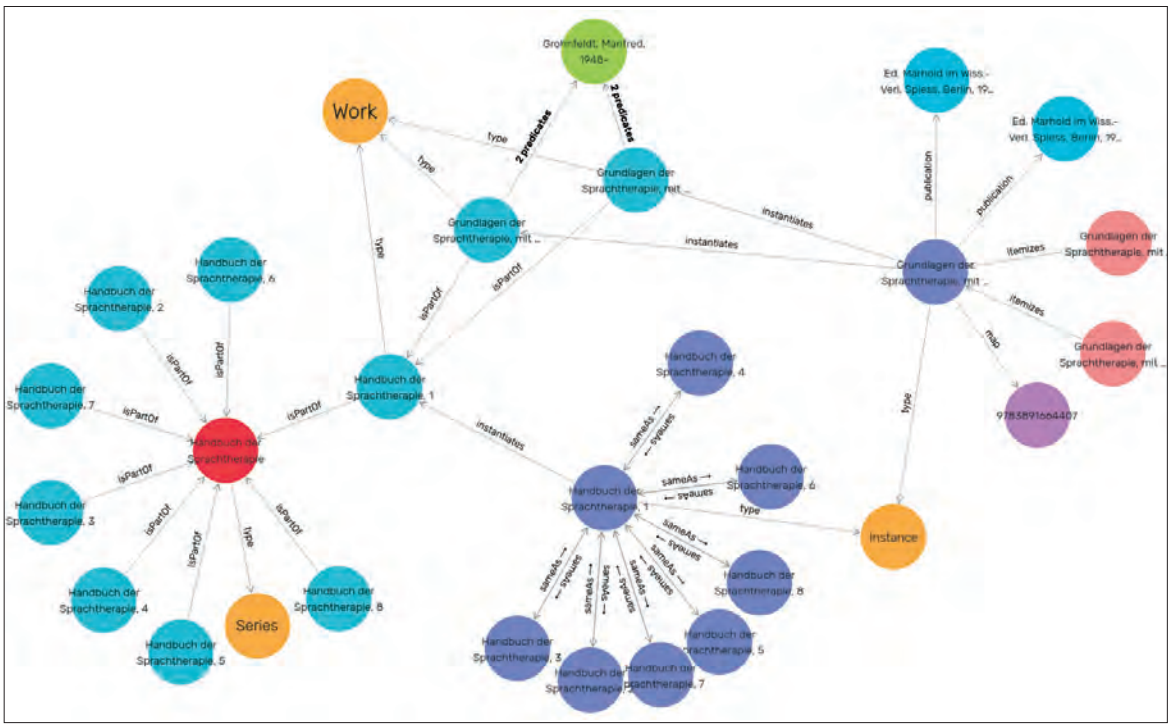

Abbildung 10: Mehrbändiges Werk

Grundsätzlich kann man bei der Analyse der Daten durchaus sehr viele nützliche Relationen erkennen. Durch geeignete SPARQL-Abfragen lassen sich Informationen aus den Daten extrahieren, die durchaus das Potential haben, das „Rauschen" in den Katalogdaten zu reduzieren.

In einer Videokonferenz mit EBSCO sowie einem von EBSCO veranstalteten „Techday" und Workshop im März 2023 konnten die Ergebnisse präsentiert werden und stießen auf großes Interesse. Offene Fragen etwa nach der konkreten Lizenz der Daten (CC0 kommt), einer Download-Funktion im Dashboard (kommt auch), einem SPARQL-Endpoint (aktuell nicht geplant) und zum Zeitplan der Einführung der „Google Syndication" in Deutschland („Henne-Ei-Problem", da Google Reichweite als Kriterium ansetzt) wurden weitestgehend beantwortet.

Es lässt sich festhalten, dass auch BiblioGraph durchaus das Potential hat, bei der Erreichung der Ziele der UB Dortmund zum Linked Open Data eine Rolle zu spielen. Jedoch zeigt sich auch hier, dass BiblioGraph nicht die alleinige Lösung ist. Zum einen liefert der Algorithmus noch zu viele unschöne Daten. Ferner ist auch hier ein nicht unerheblicher lokaler Aufwand zu betreiben, um die Daten in die Anwendung zu bekommen. Auch Effekte auf Suchen in Google sind derzeit noch nicht erkennbar. Letztlich muss hier auch berücksichtigt werden, dass BiblioGraph ein Produkt ist, welches zwar Linked Open Data erzeugt, aber trotzdem bezahlt werden will. Es lohnt sich aber, BiblioGraph weiter zu testen und auf seine Einsetzbarkeit zu prüfen.

Fazit und Ausblick

Aus den bisher betrachteten Ansätzen und Datenquellen lässt sich einiges Nachnutzbares identifizieren. So sind aus den Datenquellen, die die Daten der UB Dortmund bereits enthalten, durchaus Heuristiken ableitbar, um einen FRBR angelehnten und Werk-reduzierten Graphen zu erhalten. Auch BiblioGraph kann mit seinem Algorithmus und der „Google Syndication" durchaus punkten, ist aber als „Allheilmittel" derzeit noch nicht geeignet.

Es zeigt sich aber auch, dass es auf dem Gebiet „Linked Open Data" für das Erreichen der Ziele der UB Dortmund noch einige offene Baustellen gibt. Letztlich lässt sich feststellen, dass es ohne einen signifikanten Anteil eigener IT- und Rechenkapazitäten sowie Personalressourcen nicht geht.

Foto: Vera Münch

Abbildung 11: Eindrücke vom BiblioGraph-Techday

Hans-Georg Becker
E-Medien und Datenmanagement
Universitätsbibliothek der TU Dortmund
hans-georg.becker@tu-dortmund.de