Markus M. Geipel, Christoph Böhme, Julia Hauser,
Alexander Haffner

Herausforderung Wissensvernetzung

Impulsgebende Projekte für ein zukünftiges LOD-Konzept der Deutschen Digitalen Bibliothek

Einleitung

Ziel der Deutschen Digitalen Bibliothek[1] (DDB) ist es, das kulturelle Erbe Deutschlands, gesammelt in einer Vielzahl von Kultur- und Wissenschaftseinrichtungen, über ein Portal zugänglich zu machen und zu vernetzen. Ein institutions- und spartenübergreifender Zugang zu Kultur birgt ungeahntes Potential: Allein die Möglichkeit, mit nur einer Anfrage die Bestände zehntausender Gedächtnisorganisationen auf einmal zu durchsuchen, lässt auf neue Erkenntnisse hoffen. Angesichts solcher Chancen übersieht man jedoch leicht die Risiken, die ein zentrales Portal birgt: Wenn alles auf einen gemeinsamen Nenner herunter gebrochen wird, um in das Korsett einer einheitlichen Darstellung zu passen, wenn Verbindungen nur innerhalb dieses normierten Datensilos existieren, dann läuft das Portal Gefahr, von einem Einstiegspunkt zu einem Endpunkt zu degenerieren. Dies wäre inakzeptabel, da Wissen und Kultur von der offenen Vernetzung leben.

Das Konzept von *Linked Open Data (LOD)* formalisiert diese Vision von offenen und vernetzten Daten in Hinblick auf eine technische Umsetzung. Jedoch ist *Linked Open Data* keine fertig vorliegende Lösung, die durch die Installation einer Standardsoftware auf die Schnelle einem Web-Portal hinzugefügt werden könnte. Ein LOD-Konzept muss individuell erstellt und an die vorhandenen Daten angepasst werden. Im Falle der DDB ergeben sich zwei große Herausforderungen:

- Die Heterogenität der teilnehmenden Institutionen schlägt sich in einer immensen Datenvielfalt nieder. Wie können diese Daten spartenübergreifend modelliert und in einer gemeinsamen Ontologie verankert werden?
- In einem Datenbestand, der sich aus einer großen Zahl von Einzelbeständen zusammensetzt, existieren in vielen Fällen keine Verknüpfungen zwischen den Objekten der Einzelbestände. Solche Verknüpfungen bilden jedoch eine Grundlage von *Linked Open Data*. Wie können die Daten der DDB sowohl

1 http://www.deutsche-digitale-bibliothek.de/

innerhalb des Datenbestandes als auch mit externen Beständen verknüpft werden?

Der Umfang dieser Herausforderungen mag entmutigen. Jedoch muss das Rad nicht neu erfunden werden. Es gibt bereits eine Reihe von Projekten, von denen gelernt werden kann und deren Ergebnisse nachgenutzt werden können. Einige könnten sogar eine aktive Rolle in einem zukünftigen LOD-Konzept der Deutschen Digitalen Bibliothek spielen.

Prominent ist das europäische Gegenstück zur DDB, die Europeana. Sie steht vor denselben Herausforderungen wie die DDB und hat bereits zahlreiche Erfahrungen gesammelt, auf denen aufgebaut werden kann. Davon abgesehen wurde auch im Bibliotheksbereich bereits essenzielle Vorarbeit geleistet. Mehrere Bibliotheksportale bieten LOD sowohl für Titeldaten als auch für Normdaten an. Eine zentrale Rolle kommt hierbei dem LOD-Dienst der Deutschen Nationalbibliothek[2] zu, welcher unter anderem die Gemeinsame Normdatei in RDF (*Resource Description Framework*[3]) publiziert und damit eine institutionsübergreifende integrative Funktion erfüllt. Zahlreiche Services von Verbünden bieten Titeldaten offen in RDF an: lobid.org oder lod.b3kat.de. Projekte wie CONTENTUS[4] oder Culturegraph[5] geben wichtige Impulse im Bereich Erkennung, Abgleich und Vernetzung von Entitäten. Schließlich existieren im Bereich Normdatenvernetzung erste leicht handhabbare Standardlösungen: Zum Beispiel Dateiformate wie BEACON für den einfachen Austausch von Konkordanzlisten.

Im Folgenden werden verschiedene Projekte und Technologien vorgestellt und ihre potentiellen Beiträge zur Deutschen Digitalen Bibliothek herausgearbeitet. Wir orientieren uns dabei an zwei Themenkomplexen: Datenmodellierung und Vernetzung von Entitäten. In jedem dieser Teile wird zunächst die Problemstellung umrissen und anschließend eine Reihe von Projekten vorgestellt, welche die Aufgabenstellung auf innovative Weise adressieren. Beide Abschnitte enden jeweils mit einer Diskussion dieser Ideen in Bezug auf ihre Nutzung in der DDB. Der Artikel schließt mit einem Ausblick auf die Perspektiven der zukünftigen Entwicklung der DDB.

2 http://www.dnb.de/datendienste/linkedData
3 http://www.w3.org/RDF/
4 http://www.contentus-projekt.de/
5 http://www.culturegraph.org/

Datenmodellierung

Bibliographische Metadaten liegen in einem der etablierten Metadatenformate vor, sei es MARC 21[6], einer Variante von Pica oder dem Maschinellen Austauschformat für Bibliotheken (MAB2). Umwandlungen zwischen diesen Formaten gehören zum Tagesgeschäft großer Bibliotheken und Bibliotheksverbünde. Für die Veröffentlichung als *Linked Open Data* hat sich RDF als Format durchgesetzt. Dabei sind allerdings einige besondere Herausforderungen für die Konversion zu berücksichtigen.

Problemstellung

Den üblichen bibliographischen Metadatenformaten liegt ein anderes Datenkonzept zu Grunde als RDF. Man kann sich einen MARC-, Pica- oder MAB-Datensatz wie eine Karteikarte vorstellen, deren Felder ausgefüllt werden. RDF dagegen betrachtet Daten als ein Netzwerk. Es existieren in diesem Sinne keine einzelnen Datensätze mehr, sondern nur noch Knoten, die durch Aussagen verknüpft sind. Dieser Paradigmenwechsel macht die Umwandlung von bibliographischen Daten in RDF kompliziert. Bisher existiert keine standardisierte Konkordanztabelle zwischen einem der gängigen Bibliotheksformate und RDF. Eine solche Tabelle zu erstellen, ist auch kaum möglich, da RDF hinsichtlich des Vokabulars unbegrenzt ist: Während beispielsweise MARC 21 eine klar definierte Auswahl an Feldern vorgibt, können in RDF die „Felder" frei definiert werden: Existierende Vokabulare können kombiniert werden, oder es können sogar gänzlich neue Vokabulare entwickelt werden.

In den letzten Jahren ist in kurzer Zeit eine Vielzahl neuer Vokabulare entstanden. Dies hat – entgegen des Versprechens von *Linked Open Data* – zu einer Einschränkung der tatsächlichen Nachnutzbarkeit der RDF-Daten geführt, denn Anwendungen können nur solche RDF-Daten verarbeiten, deren Vokabulare sie kennen und verstehen. Letztendlich liegt die Prämisse beim Einsatz von Vokabularen im *Semantic Web* wie bei klassischen Metadatenformaten auf einer hohen Interoperabilität, um eine breite Nachnutzung der Daten sicherzustellen.

Die *W3C Library Linked Data Incubator Group*[7] ist gemeinsam mit Experten aus der Bibliothekscommunity und dem Bereich Semantic Web der Frage nachgegangen, wie weltweit existierende Bibliotheksdaten im Web eine höhere Inte-

6 http://www.loc.gov/marc/marcdocz.html
7 http://www.w3.org/2005/Incubator/lld/

roperabilität untereinander erlangen können. Die Gruppe schlägt hierzu zwei Ansätze in ihrem Abschlussbericht[8] vor:

1. die Nachnutzung von bereits existierenden *Linked Data*-Vokabularen,
2. die Definition von expliziten Bedeutungsbeziehungen („*Vocabulary Alignments*“) zwischen den Begrifflichkeiten in verschiedenen Vokabularen.

Nachfolgend werden Vor- und Nachteile beider Ansätze anhand potentieller Anwendungsfälle betrachtet.

Die Nachnutzung existierender Vokabulare bietet sich an, wenn es bereits Vokabulare gibt, die die zu veröffentlichenden Informationen und ihr Datenmodell vollständig oder größtenteils abdecken. Sollte das Datenmodell nicht vollständig durch das Vokabular erfasst werden, können Erweiterungsvorschläge an den Herausgeber des Vokabulars herangetragen werden. Alternativ können eigene zum Vokabular in Beziehung stehende zusätzliche Elemente spezifiziert werden.

Bei der Veröffentlichung von *Linked Open Data* im Bibliotheksumfeld zeigte sich, dass oft eine Vielzahl verschiedener Vokabulare gleichzeitig zur eindeutigen Beschreibung einer Entität (Buch, Tonträger, Person etc.) eingesetzt werden muss. Häufig ist es sogar notwendig, darüber hinaus noch zusätzlich eigene Elemente zu definieren.

Die langfristige Stabilität eines Vokabulars ist ein wichtiges Kriterium, wenn Informationen – wie bei Gedächtnisinstitutionen üblich – dauerhaft beschrieben werden sollen. Bei der Nachnutzung existierender Vokabulare verbleibt deren Definition und Pflege in der Verantwortung und dem Einfluss der jeweiligen Herausgeber. Es obliegt dem Nachnutzer, die langfristige Stabilität der Vokabulare einzuschätzen. Eine Reihe von Vokabularen hat sich als stabil bewiesen – Beispiele sind das *Dublin Core Metadata Element Set*[9] oder das *Friend-of-a-Friend (FOAF) Vocabulary*[10]. Bibliotheksspezifische Vokabulare gibt es allerdings nur wenige, die zum augenblicklichen Zeitpunkt als stabil bezeichnet werden können.

Unabhängig von der Stabilität stellt sich die Frage der Interoperabilität und Nachnutzbarkeit der bereitgestellten LOD-Repräsentation. Beim gleichzeitigen Einsatz vieler verschiedener Vokabulare muss im Blick behalten werden, ob potentielle Anwendungen mit den so beschriebenen Daten umgehen können. Versteht eine Anwendung nur einen Teil der verwendeten Vokabulare, so können die Informationen nicht oder nur teilweise interpretiert werden. Wenn beispiels-

8 http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/
9 http://dublincore.org/documents/dces/
10 http://xmlns.com/foaf/spec/

weise eine Körperschaft zum Großteil mit RDA[11]-Elementen beschrieben, die Website jedoch mit dem Element „foaf:homepage“ abgebildet wird, ist die Ressource für eine Anwendung, die nur das FOAF-Vokabular versteht, weitestgehend wertlos, da diese nur die Websiteangabe interpretieren kann, die damit verbundene Beschreibung der Körperschaft aber nicht.

Eine Kompromisslösung wäre die Beschränkung auf nur ein Vokabular, das eine partielle Abbildung des Datenmodells erlaubt, insoweit es mit dem präferierten Vokabular möglich ist. Einen solchen minimalen Ausschnitt darzustellen, ist jedoch meistens nicht im Interesse der veröffentlichenden Organisation.

Wenn für die vollständige Abbildung des eigenen Datenmodells zu viele verschiedene Vokabulare zusammen getragen werden müssen, besteht die Alternative, ein komplett eigenes Vokabular zu definieren. Der sich daraus ergebende Nachteil ist, dass man für die Pflege und die Verbreitung bzw. Bekanntmachung selbst verantwortlich ist. Ohne eine Abbildung auf bestehende Vokabulare kann ein solches selbst definiertes Vokabular nicht ohne weiteres von anderen Anwendungen interpretiert werden.[12]

Laut der Empfehlungen der *Library Linked Data Incubator Group* ist die Spezifikation eines neuen Vokabulars dementsprechend nicht ausreichend. Der Herausgeber des Vokabulars ist zusätzlich in der Pflicht, Elemente – wo möglich – mit Elementen aus existierenden Vokabularen zu verknüpfen („*Vocabulary Alignment*“). Im Optimalfall sollte ein *Alignment* nicht nur mit einem Vokabular, sondern mit mehreren vorgenommen werden. Am Beispiel der Körperschaft wird der Vorteil deutlich: Wenn für den Namen der Körperschaft ein eigenes Element verwendet wird und diesem ein Äquivalent sowohl in den *RDA Element Sets* als auch im FOAF-Vokabular zugewiesen ist, können Anwendungen, die eines der beiden Vokabulare verstehen, über *Reasoning* auch das neu definierte Element interpretieren.

Für das *Alignment* gibt es neben der Ausweisung der Äquivalenz zweier Elemente auch weitere Konzepte wie Inverse oder Unterklasse, wodurch eine detailliertere Ausweisung der Beziehungen zwischen Vokabularen möglich ist.

Der Ansatz des *Vocabulary Alignments* ist relativ neu und daher in der Praxis noch nicht weit verbreitet[13]. Die Entscheidung, welcher Ansatz der besser geeignete ist, ist immer abhängig von den bereitzustellenden Daten. Die im Folgenden

11 http://rdvocab.info/

12 Vgl. auch das in diesem Band von Klee vorgestellte 5-Sterne-Schema zur Veröffentlichung von Vokabularen.

13 www.kimforum.org/Subsites/kim/SharedDocs/Downloads/DE/Berichte/internationalisierungDerGndDurchDasSemanticWeb.html

beschriebenen Anwendungsfälle sollen als Hilfestellung für die Entscheidungsfindung dienen.

Impulsgebende Projekte

Für eine LOD Modellierung kann die DDB bereits auf die Erfahrungen zahlreicher Projekte zurückgreifen. Im Folgenden ein Überblick.

Europeana

Ziel der Europeana ist es, Europas wissenschaftliches und kulturelles Erbe online zugänglich zu machen. Europeana ist daher das europäische Gegenstück zur DDB und steht in der LOD-Modellierung daher vor den gleichen Herausforderungen. Nachdem Europeana nach der Entwicklung ihres ersten Datenmodells, den *Europeana Semantic Elements (ESE)* festgestellt hatte, dass durch ein minimalistisches und flaches *Dublin Core* ähnliches Modell die Semantik der Original-Metadaten verloren geht und keine Verknüpfung mehr möglich ist, entschloss man sich zur Entwicklung eines neuen Datenmodells, dem *Europeana Data Model (EDM)*[14].

Das EDM erlaubt eine Unterscheidung zwischen dem realen Objekt (z.B. Buch, Bild, Akte, mediale Aufzeichnung), seiner digitalen Repräsentation und den es beschreibenden Metadaten. Durch die Bereitstellung von Metadaten durch verschiedene *Data Provider* müssen in Europeana mehrere Sichten auf dasselbe Objekt möglich sein, selbst mit gegebenenfalls einander widersprechenden Aussagen[15].

Für die technische Umsetzung nutzt das EDM Elemente etablierter Vokabulare: *Dublin Core* beschreibt die deskriptiven Metadaten, OAI-ORE wird für die Abbildung der Aggregationen (Proxy-Konzept) von realen Objekten und Web-Ressourcen verwendet, die diese realen Objekte repräsentieren, und das *Simple Knowledge Organization System (SKOS)*[16] erfasst die Repräsentation von Wissensorganisationsystemen und deren Verbindungen untereinander. Um das zugrundeliegende Datenmodell abzubilden, wurden darüber hinaus eigene Klassen und Eigenschaften spezifiziert, die nicht durch existierende Vokabulare abgedeckt werden konnten.

14 http://pro.europeana.eu/edm-documentation
15 http://pro.europeana.eu/web/guest/tech-details
16 http://www.w3.org/2004/02/skos/

Das EDM setzt ebenfalls auf *Vocabulary Alignment*, z.B. zu der CIDOC CRM Repräsentation von FORTH-ICS[17], um die Interoperabilität zu erhöhen.

Erlangen CRM/OWL

Das Erlangen CRM/OWL[18] ist eine OWL-DL 1.0 Implementierung des *CIDOC Conceptual Reference Model (CIDOC CRM)*. Im Rahmen der DDB ist dieses Projekt relevant, da das Internformat der DDB auf CRM aufsetzt und Erlangen CRM/OWL die bisher einzige vollständige Umsetzung von CRM in einer Ontologie darstellt. Sollte die DDB weiter auf CRM setzen, mag Erlangen CRM/OWL als Vorlage dienen.

Die Ontologie wurde gemeinschaftlich vom Lehrstuhl für Künstliche Intelligenz der Friedrich-Alexander-Universität Erlangen-Nürnberg, dem Germanischen Nationalmuseum und dem Zoologischen Forschungsmuseum Alexander Koenig erstellt. Sie findet derzeit Einsatz im Projekt Wissenschaftliche Kommunikationsinfrastruktur (WissKI)[19] und im Projekt *ArcheoInf*[20].

Das CIDOC CRM ist in ISO 21127:2006[21] standardisiert und stellt ein formalisiertes Begriffsmodell bereit, um die Integration, Zugriffsvermittlung und den Austausch verschiedenartig strukturierter Informationen aus dem Bereich des kulturellen Erbes zu unterstützen. Das objektorientierte Modell wird seit den 1990er Jahren entwickelt und basiert auf in der Praxis existierenden Datenstrukturen. Das Schlüsselkonzept des Modells ist die explizite Ereignismodellierung. Die im Dezember 2011 veröffentlichte Version 5.0.4 umfasst 90 Klassen in einer Klassenhierarchie und 149 eindeutige Eigenschaften[22].

Das Erlangen CRM/OWL ist eine Interpretation des CIDOC. Aktuell ist die Ontologie die einzige aktiv gepflegte Umsetzung des CIDOC CRM in OWL. Die Implementierung setzt alle Merkmale des CIDOC CRM um. Zusätzlich existieren Erweiterungen mit einer FRBRoo[23]-Umsetzung sowie einer *Time-Spans*-Umsetzung in RDF/OWL.

Im Erlangen CRM/OWL werden keine bereits existierenden Vokabulare nachgenutzt und es wird kein *Vocabulary Alignment* vorgenommen.

---

17 http://www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.2_english_label.rdfs
18 http://erlangen-crm.org/current-version
19 http://wiss-ki.eu/
20 Vgl. den Beitrag von Lins/Becker in diesem Band
21 http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=34424
22 http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf
23 http://www.cidoc-crm.org/frbr_inro.html

Linked Data Service für deutschsprachige Normdaten

In der DDB werden Normdaten als verbindendes Element zwischen verschiedenen Objekten eine zentrale Rolle spielen. Im Idealfall bilden sie das verbindende Element zwischen den verschiedenen Sammlungen. Normdaten sind damit wichtiger Bestandteil eines Linked Data Konzepts für die DDB und mit dem *Linked Data Service* der Deutschen Nationalbibliothek (DNB) steht bereits eine etablierte Normdatenmodellierung als Vorlage zur Verfügung. Seit April 2012 wird eine vollständige Abbildung der Gemeinsamen Normdatei (GND) veröffentlicht. Das bereitgestellte Datenset umfasst Ressourcenbeschreibungen für individualisierte und nicht-individualisierte Personen, Familien, Körperschaften, Kongresse und Veranstaltungen, Geografika, Schlagwörter und Werke.

Für die *Linked Data*-Repräsentation konnte auf die objektorientierte Struktur der GND zurückgegriffen werden. Dadurch gestaltete sich die Datenanalyse relativ unkompliziert, da die enthaltenen Entitäten bereits eindeutig definiert und sogar in der klassischen MARC 21-Abbildung differenziert werden. Durch weitere Analysen des Erfassungsleitfadens sowie des Datenformates konnte ein Datenmodell abgeleitet werden.

Die Mächtigkeit der GND konnte durch kein verfügbares Vokabular abgedeckt werden. Selbst beim Einsatz von Elementen verschiedener Vokabulare hätte für die erstrebte Abbildung ein Großteil eigener Elementdefinitionen vorgenommen werden müssen. Aus diesem Grund wurde eine komplett eigene GND Ontologie[24] definiert.

Entsprechend den Empfehlungen der *W3C Library Linked Data Incubator Group* wurde für die GND Ontologie – als eigenständiges Vokabular – der Ansatz des *Vocabulary Alignment* eingesetzt. In einem ersten Schritt wurden FOAF und die *RDA Element Sets*[25] verknüpft. *Alignments* zu weiteren Vokabularen sollen folgen.

Die auf OWL basierende Ontologie enthält 49 Klassen in einer Klassenhierarchie, 162 Objekteigenschaften, 54 Datentypeigenschaften und eine Annotationseigenschaft für die Ausweisung der MARC 21-Äquivalenz. Insgesamt konnten 14 *Alignments* zu FOAF und 90 zu den *RDA Element Sets* vorgenommen werden.

24 http://d-nb.info/standards/elementset/gnd#
25 http://rdvocab.info/

Linked Data Services für Titeldaten

In der Bibliothekswelt steht eine wachsende Anzahl von LOD-Diensten zur Verfügung, denn in den letzten Jahren haben immer mehr Bibliotheken erkannt, welches Potential sich mit der Veröffentlichung ihrer Daten entfaltet: Die qualitativ hochwertigen, größtenteils intellektuell gepflegten bibliographischen Daten, die bisher in den Bibliothekskatalogen eingekapselt waren, werden neuen Nutzergruppen zugänglich und können Grundlage für die Etablierung neuer Services und Anwendungen werden. Bibliographien werden zunehmend als Teil des *World Wide Web* begriffen und der daraus entstehende Mehrwert wird genutzt (Kett 2012).Die LOD-Dienste für Titeldaten können daher als Inspirationsquelle für zukünftige LOD-Dienste der DDB dienen.

Die Präsentation von Titeldaten in der *Linked Data Cloud* ist allerdings heterogen: Sowohl einzelne Bibliotheken (z.B. UB Mannheim[26] und DNB[27]) als auch Bibliotheksverbünde (z.B. hbz[28], BVB und KOBV[29] und HeBIS[30]) entwickeln und betreiben *Linked Data Services*. Da die Dienste unabhängig voneinander entstanden sind, die Institutionen teilweise verschiedene Katalogisierungsrichtlinien haben und mit unterschiedlichen Bibliothekssystemen und Formaten arbeiten, sind die Datenmodellierungen und technischen Umsetzungen vielfältig. Bisher verwenden allerdings alle den Ansatz der Nachnutzung bestehender Vokabulare. Teilweise wurden eigene Elemente geprägt, wenn bestehende Ontologien keine geeigneten Elemente boten. Sowohl der Umfang der umgesetzten Quellformate als auch die Auswahl der jeweiligen RDF-Elemente und deren Einsatz variieren.

Die Varianz der eingesetzten Vokabulare und des Umfangs der Umsetzung machen eine Nachnutzung der bibliographischen Daten schwierig. Den *Linked Data Services* steht jedoch ein gemeinsames Ziel voran: Bibliographische Daten sichtbar und für Bereiche außerhalb des klassischen Bibliothekswesens zugänglich und nutzbar zu machen. Um die Interoperabilität zwischen unterschiedlichen bibliographischen Datenquellen zu gewährleisten, sollten diese Dienste also möglichst eine gemeinsame Sprache sprechen.

Viele *Linked Data Services* befinden sich noch in einem experimentellen Stadium und werden weiter ausgearbeitet. In dieser Phase ist die Datenmodellierung noch nicht abschließend geklärt. Deshalb ist es sinnvoll, Kompetenzen und bereits gesammelte Erfahrungen zu bündeln. Daher hat sich eine deutschland-

26 http://data.bib.uni-mannheim.de/
27 http://www.dnb.de/datendienste/linkedData
28 http://lobid.org/
29 http://lod.b3kat.de/
30 http://www.hebis.de/de/1ueber_uns/projekte/lod/lod_index.php

weite Arbeitsgruppe gegründet, die eine Harmonisierung der Titeldatenmodellierungen anstrebt[31]. Hierin vertreten sind Mitarbeiter aller deutschen Bibliotheksverbünde, der Deutschen Nationalbibliothek sowie einige weitere interessierte und engagierte Kollegen mit entsprechender Expertise. Die Arbeitsgruppe agiert seit April 2012 als eine Untergruppe der DINI-AG KIM[32].

In der Arbeitsgruppe wird zunächst unabhängig vom Ausgangsformat aus fachlicher Sicht analysiert, welche Metadaten zur Beschreibung eines bibliographischen Datensatzes erforderlich sind und wie sie logisch gruppiert werden können. Hierbei darf nicht aus den Augen verloren werden, dass die Zielgruppe im Zusammenhang mit *Linked Data* weit über die Bibliothekswelt hinausgeht. Auch die Anforderungen neuer Zielgruppen müssen also verstanden werden und in die Datenauswahl und -modellierung mit einfließen. Erst wenn diese grundsätzlichen und konzeptionellen Fragen geklärt sind, können konkrete *Mappings* erarbeitet werden. Dabei bietet sich MARC 21 als gemeinsames einheitliches Ausgangsformat an.

Auch international gibt es Bestrebungen, eine Harmonisierung der Titeldatenmodellierungen herbeizuführen. Hierbei ist die *Bibliographic Metadata Task Group*[33] der *Dublin Core Metadata Initiative* federführend. Die Zusammenarbeit und Vernetzung ist hier äußerst wichtig, um parallele Entwicklungen zu beobachten und Synergieeffekte nutzen zu können. So wird auch die DINI-AG-KIM Titeldaten-Gruppe weiter verfolgen, welche Empfehlungen auf internationaler Ebene erarbeitet werden und eigene Impulse aus der nationalen Kooperation mit einbringen. Ziel der nationalen Arbeitsgruppe ist es, einen *Best Practice Guide* zu erarbeiten und so einen Quasistandard im Bereich bibliographischer Metadaten zu etablieren, der die Grundlage der Modellierung bibliographischer Daten in der Deutschen Digitalen Bibliothek sein könnte.

Diskussion

Insgesamt zeigt sich ein deutlicher Trend zur Standardisierung und Abstimmung im Bereich Ontologien. Im Bibliothekswesen sind beispielsweise die o.g. Arbeitsgruppen entstanden. Im Kulturbereich allgemein ist die Konvergenz von Ontologien und Modellen noch weniger stark ausgeprägt, jedoch zeichnet sich eine Tendenz zu EDM beziehungsweise CIDOC CRM ab.

31 https://wiki.d-nb.de/display/DINIAGKIM/Titeldaten+Gruppe
32 http://www.dini.de/ag/standards/
33 http://wiki.dublincore.org/index.php/Bibliographic_Metadata_Task_Group

In Hinblick auf ein LOD Konzept für die DDB sollte daher auf Vorhandenem aufgebaut werden. Nicht zuletzt, da die DDB als Aggregator für Europeana agieren soll, ist Kompatibilität mit den sich etablierenden Standards geboten. Auf ein einziges Format zu setzen, erspart Mehraufwand und minimiert Umwandlungsverluste, die im Zuge von Formattransformationen kaum vermeidbar sind. Mit Europeana liegt also eine Orientierung an EDM auf der Hand, und sei es nur in Form von *Vocabulary Alignments*.

Was die Normdaten betrifft stehen der DDB zwei Wege offen: erstens selbst Normdaten zu halten und diese als Linked Data anzubieten. In diesem Fall kann der Linked Data Dienst der DNB und insbesondere die Ontologie der GND als Blaupause dienen. Der zweite Weg wäre existierende Normdaten wie die GND zu nutzen und zu referenzieren. Für die Darstellung im DDB Portal könnten die Daten live über den Linked Data Dienst der DNB oder entsprechender Normdatendienste weiterer Institutionen abgerufen werden. Dieser letztere Weg entspräche dem Vernetzungsgedanken von LOD und des Semantic Web am besten.

Entitäten vernetzen

Die DDB wird eine große Anzahl von Sammlungen in sich vereinen. Sollen diese mehr sein als nur die Summe der Einzelteile, müssen neue Verknüpfungen hergestellt werden, die die Sammlungen zu einem großen Ganzen verschmelzen, denn Wissen lebt von Verknüpfungen, und neues Wissen entsteht durch das Herstellen von Bezügen. Betrachten wir folgende Aussagen:

1. Charles Lutwidge Dodgson (1832 - 1898) war ein englischer Mathematiker und Photograph. Er unterrichtete am Christ Church College in Oxford.
2. Lewis Carroll ist der Autor von „Alice in Wonderland“.

Fügen wir nun die Verknüpfung ein, dass Lewis Carroll und Charles Lutwidge Dodgson ein und dieselbe Person sind, gewinnen wir Wissen, das über diese eine Verknüpfung hinausgeht: Wir können jetzt den Entstehungszeitraum von „Alice in Wonderland“ eingrenzen, da wir nun die Lebensdaten von Carroll kennen. Wir wissen außerdem, dass der Buchautor Carroll Engländer war, dass er Mathematiker war, und wir wissen außerdem, dass Charles Lutwidge Dodgson Pseudonyme benutzt hat. Die Verknüpfung zwischen den beiden Namen bietet so auch in sich schon einen Mehrwert. Ein verknüpfter Datenbestand ist mehr als nur die Summe der Teildatenbestände.

Problemstellung

In der Praxis ist es leider oft sehr schwer, Verknüpfungen herzustellen. Dies hat mehrere Gründe: Zum einen wurden Datenbestände meist nicht in Hinblick auf eine externe Verknüpfung angelegt. Es existieren daher meist nur interne Verknüpfungen, die über IDs oder Schlüssel hergestellt werden, die nur innerhalb des spezifischen Datenbestands Bedeutung und Eindeutigkeit besitzen. Oft jedoch sind nicht einmal interne Verknüpfungen gegeben, insbesondere bei Daten aus vordigitaler Zeit. Man stelle sich alte Katalogkarten vor: Auf ihnen sind keine IDs für Autoren oder Schlagworte verzeichnet, lediglich Namen, also Zeichenketten ohne Anspruch auf Eindeutigkeit. Das war in jener Zeit sinnvoll, in der die Katalogkarten von Menschen und nicht von Maschinen benutzt wurden. Für die heutigen Anforderungen im *Semantic Web* sind Namen jedoch ein ernsthaftes Problem. Der Kern dieses Problems wird im normalen Sprachgebrauch kaschiert: Man sagt zum Beispiel ein Buch handle von „Michael Jackson". Tatsächlich ist die Aussage aber: Das Buch handelt von einer Person *X*, und *X* hat den Namen „Michael Jackson". Dieses *X* kann tatsächlich für eine Vielzahl von Personen stehen: Eben all jene, die den Namen „Michael Jackson" tragen. Die Mehrheit der Menschen wird bei „Michael Jackson" unmittelbar an den Musiker denken. Ein Whiskeykenner jedoch würde für jene Person *X* mit Namen „Michael Jackson" den britischen Whiskey Degustationsexperten Michael Jackson einsetzen[34], ein US-Historiker einen General und ein Einwohner Dublins einen Bischof.

Wenn vernetzt werden soll, kann also nicht alleine mit Namen gearbeitet werden, es sind Entitäten notwendig, die durch eindeutige Identifikationen bezeichnet werden. Die Entitäten ihrerseits müssen beschreibende Attribute erhalten und als eigene Datenbasis gepflegt werden. Zwei Unterprobleme ergeben sich in Bezug auf Entitäten: Erstens das Problem, zwei Entitäten als gleich oder in Bezug zueinander stehend zu erkennen, wie im Falle von Lewis Carroll und Charles Lutwidge Dodgson. Zweitens, das komplementär aufgestellte Problem gleich anmutende Entitäten auseinander zu dividieren, wie im Falle der zahlreichen Michael Jacksons.

Impulsgebende Projekte

Im Folgenden werden Projekte vorgestellt, in denen Daten verknüpft werden. Unterschieden werden dabei Projekte, die auf den menschlichen Intellekt setzten

34 http://d-nb.info/gnd/120744228

und Projekte, die Verknüpfungen mithilfe von Algorithmen erzeugen. Beide Herangehensweisen könnten in der DDB zu Einsatz kommen.

Intellektuelle Verknüpfung

Auf menschliche Arbeitskraft wird zurückgegriffen, wenn die Entitäten sehr wenig Kontext mit sich führen. Die Verknüpfung zwischen Lewis Carroll und Charles Lutwidge Dodgson ließe sich mit den im Beispiel angeführten Informationen nur schwerlich automatisch erzeugen. Eine intellektuelle Verknüpfung ist allerdings nur gangbar, wenn die Datenmenge begrenzt ist oder sehr viele freiwillige Helfer zur Verfügung stehen wie etwa in *Crowdsourcing*-Ansätzen[35]. Das emblematische Beispiel hierfür ist Wikipedia.

In der deutschsprachigen Wikipedia werden Artikel zu Personen seit 2005 mit externen Datenquellen verknüpft[36]. In einem Kooperationsprojekt zwischen Wikipedia und der Deutschen Nationalbibliothek wurden zunächst Einträge zu Personen in der Wikipedia von Wikipedia-Mitarbeitern intellektuell mit den dazugehörigen Einträgen in der Personennormdatei (PND)[37] verknüpft (Danowski 2007). Seit 2009[38] werden zusätzlich auch Verknüpfungen zu den Personendatenbanken der *Library of Congress*, der japanischen Nationalen Parlamentsbibliothek und dem *Virtual International Authority File (VIAF)* erfasst[39]. Seit der Einführung der Gemeinsamen Normdatei (GND) werden in der Wikipedia GND-Nummern anstelle von PND-Nummern erfasst.

Zur strukturierten Erfassung der verschiedenen Verknüpfungen in der Wikipedia wird eine Vorlage bereitgestellt[40], in die die Wikipedia-Autoren die Referenzen der verschiedenen Normdateien eintragen können. Durch die Verwendung einer Vorlage wird die Erstellung von Verknüpfungen erleichtert und ein einfacher maschineller Zugang zu den Normdatenverknüpfungen sichergestellt. Dadurch wird es zum einen möglich, die Normdatenverknüpfungen in den Wikipedia-Artikeln in einer standardisierten Form anzuzeigen und zum anderen können die Verknüpfungen so leicht automatisiert extrahiert werden, um bei-

35 *Crowdsourcing* setzt auf die Beteiligung einer großen Anzahl von Freiwilligen, welche eine große Aufgabe durch viele kleine Beiträge lösen. Ein klassisches Beispiel ist Wikipedia.

36 http://de.wikipedia.org/w/index.php?title=Hilfe:Personendaten&oldid=105294857#Zur_Geschichte_der_Personendaten

37 Die Personennormdatei ist mittlerweile in der Gemeinsamen Normdatei aufgegangen.

38 http://de.wikipedia.org/w/index.php?title=Hilfe:Personendaten&oldid=105294857#Zur_Geschichte_der_Personendaten

39 http://de.wikipedia.org/w/index.php?title=Vorlage:Normdaten&stableid=105715515

40 Ebd.

spielsweise Konkordanzen zwischen der Wikipedia und den verknüpften Normdatenquellen zu erstellen. Wikipedia stellt eine entsprechende Konkordanz im BEACON-Format[41] bereit[42].

Zur Unterstützung bei der Verknüpfung wurde ein Tool entwickelt, das anhand eines Wikipedia-Artikels automatisch passende Einträge in der Personennormdatei sucht. Der Benutzer muss dann nur noch bestätigen, dass der vorgeschlagene Eintrag korrekt ist, um automatisch die Normdaten-Vorlage für den Wikipedia-Artikel auszufüllen (Danowski 2007).

Der Fokus der Normdatenverknüpfung in der Wikipedia lag bisher bei Personen. Allerdings unterstützt die „Normdaten"-Vorlage auch die Erfassung von Verknüpfungen anderer Entitätstypen wie etwa Geografika, Körperschaften oder Sachbegriffe.

Algorithmische Verknüpfungen

Algorithmen können mit entsprechenden Kontextinformationen wie Publikationslisten, biographischen Daten oder Identifikationsnummern erstaunlich gute Ergebnisse bei der Verknüpfung kultureller Objekte liefern. Im Folgenden werden drei Projekte vorgestellt, die algorithmisch Verknüpfungen erzeugen.

CONTENTUS

Im CONTENTUS-Projekt wurden neue Technologien für semantisch erschlossene, multimediale Archive erforscht und entwickelt. Neben Schwerpunkten in der automatischen Erschließung verschiedener Medientypen spielte auch die Verknüpfung der dabei gewonnenen Metadaten untereinander und mit externen Datenquellen eine wichtige Rolle. Ziel dieser Verknüpfung war die Integration der verschiedenen Archivobjekte in einem gemeinsamen Wissensnetz.

Als zentrale Komponente, um dieses Ziel zu erreichen, wurde im Projektverlauf die Nutzung von Entitäten identifiziert: Es zeigte sich im Projekt, dass zwischen einzelnen Archivobjekten selten ein direkter Zusammenhang besteht, sondern dass eine Verbindung zwischen zwei Objekten meist über eine Entität – etwa einen Ort oder eine Person – entsteht. Aus diesem Grund lag ein Fokus der Arbeiten im CONTENTUS-Projekt auf der Generierung von Metadaten, die Entitäten eindeutig referenzieren. Daneben sollten die von den Archivobjekten referenzierten Entitäten aber auch mit externen Datenquellen verknüpft werden, um auf diese Weise zusätzliche Informationen im Wissensnetz zugänglich zu machen.

41 http://de.wikipedia.org/w/index.php?title=Wikipedia:BEACON&oldid=105449989
42 PND-Beacon-Datei: http://toolserver.org/~apper/dewp_pnd_beacon.txt

In CONTENTUS wurde ein automatisierter Abgleich zwischen den Normdateien der Deutschen Nationalbibliothek und Wikipedia für die in den Archivobjekten gefundenen Entitäten angestrebt. Während dies für Personen mit Hilfe der intellektuell durchgeführten Verknüpfung leicht möglich war, musste für Geografika ein eigenes Verfahren entwickelt werden. Dieses Verfahren basierte auf einem Vergleich der Ortsnamen sowie weiterer Charakteristika wie Homonymzusätzen („Frankfurt am Main" oder „Frankfurt/Oder") und übergeordneten geografischen Einheiten.

Unter Verwendung dieses Verfahrens konnten geographische Entitäten erfolgreich mit der Schlagwortnormdatei[43] und der deutschen Wikipedia verknüpft werden.

Culturegraph

Erklärtes Ziel der Plattform Culturegraph ist es, eine einheitliche, verlässliche und persistente Referenzierbarkeit von kulturellen Erzeugnissen zu ermöglichen. Dazu wurden im ersten Schritt in einem gemeinsamen Projekt der Deutschen Nationalbibliothek (DNB) und des Hochschulbibliothekszentrums des Landes Nordrhein-Westfalen (hbz) die Kataloge der deutschen Bibliotheksverbünde abgeglichen und statistisch ausgewertet. Die Ergebnisse sind als *Linked Open Data* abrufbar. In weiteren Projekten wird die Vernetzung von Normdaten vorangetrieben.

So werden beispielsweise im Projekt *Culturegraph Authorities* Rückverknüpfungen aus den Normdaten zu den Titeldatensätzen errechnet. In der Praxis kann dadurch der Normdatensatz als Sucheinstieg genutzt werden: Von einem Normdatensatz in Culturegraph – zum Beispiel „Bertold Brecht" – führen Verknüpfungen zu allen deutschen Bibliotheksverbünden sowie einer Zahl weiterer Institutionen, welche Medieneinheiten von Bertold Brecht vorhalten. Diese Verknüpfungen zu weiteren Institutionen wurden aus frei zugänglichen BEACON-Dateien[44] gelesen. BEACON-Dateien verzeichnen Konkordanzen von einem Datenbestand zur GND und werden in der Regel von der Institution erstellt und gepflegt, in deren Verantwortung auch der entsprechende Datenbestand liegt. Culturegraph spielt in diesem Fall also die Rolle des Integrators, welcher verteilt verfügbare Verknüpfungsinformationen sammelt und zentral anbietet.

In weiteren Entwicklungsschritten sollen Konkordanzen zu weiteren Datenbeständen regelmäßig errechnet und zur Datenanreicherung genutzt werden. Diese Anreicherung legt den Grundstein für neue intelligente Suchverfahren sowie Verfahren zur Qualitätssicherung.

43 Die Schlagwortnormdatei ist mittlerweile in der Gemeinsamen Normdatei aufgegangen.
44 http://de.wikipedia.org/wiki/Wikipedia:BEACON

Culturegraph setzt auf rein algorithmische Verknüpfungen. Daten werden auf einem Hadoop-Cluster[45] verarbeitet und anschließend auf www.culturegraph.org veröffentlicht. Die in diesem Rahmen entwickelte Software ist als Open Source veröffentlicht[46]. Damit stellt Culturegraph. nicht nur eine Webplattform dar, sondern auch eine frei verfügbare Werkzeugpalette für Metadatenverarbeitung und Vernetzung. Eine Nachnutzung auch im DDB-Kontext ist daher in vielerlei Hinsicht möglich.

VIAF

VIAF (Virtual International Authority File) ist ein Gemeinschaftsprojekt mehrerer Nationalbibliotheken, welches von dem weltweit tätigen Bibliotheksdienstleister OCLC technisch umgesetzt wird. Ziel ist die Errechnung einer Konkordanz zwischen den Personennormdateien der teilnehmenden Institutionen. Der Berechnung liegen dabei mit der entsprechenden Person verknüpfte Titeldaten zu Grunde. Neuberechnungen finden in regelmäßigen Intervallen statt. Zeitliche Stabilität der so erzeugten Personenidentifikationsnummern (VIAF Nummern) wird garantiert, indem die Information über ehemalige Konkordanzen weiterhin gespeichert wird und entsprechende Anfragen ausgelöst werden.

Diskussion

Betrachtet man die vorgestellte Projektauswahl, fällt auf, dass sich insbesondere die algorithmische Verknüpfungserstellung als eigenständige Tätigkeit etabliert, die nicht an die Urheber oder Lieferanten der Daten gebunden ist. Weder CONTENTUS noch VIAF noch Culturegraph.org sind Urheber von Daten in dem Sinne, in dem beispielsweise die Bibliotheken oder Bibliotheksverbünde Urheber von Metadaten sind. Diese Projekte und Dienste sind also vielmehr als Mehrwertdienste aufzufassen.

Auch die DDB ist selbst kein Urheber von Metadaten und die an die DDB gelieferten Daten werden in zunehmendem Maße bereits von den entsprechenden Lieferanten als LOD veröffentlicht. Welchen Mehrwert kann die DDB mit einem LOD-Dienst also liefern?

Ein wichtiger Aspekt ist der spartenübergreifende Ansatz der DDB. In diesem Sinne könnte die DDB als LOD-Knotenpunkt die LOD-Dienste der eigentlichen Metadatenurheber miteinander verbinden. In diesem Falle könnte sich die DDB

45 Open Source Software für das verteilte Verarbeiten großer Datenmengen auf einem Rechnercluster. Siehe http://hadoop.apache.org/

46 http://github.com/culturegraph

auf die Verknüpfungen konzentrieren und die Auslieferung der eigentlichen Daten den Urhebereinrichtungen überlassen. Ob dies durch Algorithmen erreicht wird oder durch Ansätze, die sich mehr an Projekten wie Wikipedia orientieren, bleibt zu erörtern. Angesichts der vielen Vernetzungsprojekte ist auch die gezielte Nachnutzung und Anpassung bereits vorhandener Verknüpfungsdaten eine Option, beispielsweise in Form von BEACON-Dateien, Anfragen an Culturegraph oder der Nutzung von Datenabzügen von VIAF.

Ausblick und zukünftige Herausforderungen

In diesem Artikel wurde eine Reihe von Projekten unter den Aspekten der Datenmodellierung und der Entitätsverknüpfung vorgestellt. Motivation war es, Impulse zu sammeln und die Koordinaten eines zukünftigen LOD-Konzeptes der Deutschen Digitalen Bibliothek abzustecken.

Es wurde aufgezeigt, dass es in zahlreichen Projekten bereits ein breites Spektrum von Antworten auf die Herausforderungen von *Linked Open Data* gibt. Auf vieles kann aufgebaut werden. Neben diesen Erkenntnissen bleiben jedoch auch offene Fragen. Fragen, die sich in den vorgestellten Projekten bisher nur schwach abzeichnen, jedoch mittelfristig an Bedeutung gewinnen werden.

Da ist zum einen die Frage der Provenienz von Daten[47]. Je mehr Daten gemischt und weiterverarbeitet werden, desto deutlicher stellt sich die Frage, wie Herkunft und eventuell mit den Daten verknüpfte Rechte mitgeführt werden können. Europeana löst diese Herausforderung in Bezug auf die Rechtefrage, indem nur Daten mit einer CC0-Lizenz angenommen werden. Daten also, für die „keine Rechte vorbehalten“ werden. Dies umgeht die Lizenzfrage elegant, löst aber nicht das Problem der Datenherkunft. Daher ist es nicht immer ein geeigneter Ansatz. Sobald die tatsächliche Herkunft von Daten relevant wird, etwa um ihre Vertrauenswürdigkeit zu bestimmen, stößt die Lösung der Europeana an ihre Grenzen und die Frage der Datenherkunft kann nicht mehr umgangen werden.

Eine weitere Herausforderung wird die Orientierung am Endnutzer sein. Welche nützlichen LOD-Anwendungen können mit den Daten der DDB realisiert werden? Eine RDF- oder SPARQL-Schnittstelle ist lediglich eine Voraussetzung für neue innovative Anwendung, aber nicht das endgültige Ziel. Die Anwendungen selbst warten noch darauf, geschrieben zu werden. Dies wird im Bereich LOD sowohl die größte als auch spannendste Herausforderung sein. Die reichhaltigen

47 Mit dem Thema Provenienz beschäftigt sich im Detail der Artikel „Die Provenienz von Linked Data“ im vorliegenden Sammelband.

Erfahrungen aus den hier beschriebenen Projekten sollen dabei helfen, das Rad nicht neu erfinden zu müssen, sondern alle Energien auf die neuen Herausforderungen zu konzentrieren.

Literaturverzeichnis

Danowski, Patrick; Pfeifer, Barbara: Wikipedia und Normdateien: Wege der Vernetzung am Beispiel der Kooperation mit der Personennormdatei. In: Bibliothek Forschung und Praxis 31 (2007), Nr. 2, S. 149–156; DOI: 10.1515/BFUP.2007.149.

Kett, Jürgen; Manecke, Mathias; Beyer, Sarah: Die Nationalbibliografie im Zeitalter des Internets. In: ZfBB 59(2012), 2, S. 72

Danksagung

Unser besonderer Dank gilt Herrn Dr. Jan Hannemann (Deutsche Nationalbibliothek) für seine umfangreichen Kommentare und Anregungen.