DIENSTLEISTUNG DER ZUKUNFT: LINKED OPEN DATA

Als wichtigste Plattform zur Publikation und Rezeption von Informationen hat sich in den letzten beiden Jahrzehnten das *World Wide Web* entwickelt. In ihm finden sich Dokumente unterschiedlichster Medientypen, die wiederum auf andere Dokumente verweisen. Maschinenlesbare Rohdaten sind dort eher selten zu finden, da im dort vorherrschenden Dokumentenmodell hauptsächlich Textstrukturen annotiert werden; eine Beschreibung reiner Daten wird nur sehr rudimentär unterstützt. In Bibliotheken finden sich Daten in unterschiedlichsten Anwendungen: Zuerst denkt man natürlich an Metadaten im Bibliothekskatalog. Ein Titeldatensatz besteht dabei aus Normdaten zu Personen, Körperschaften und Schlagworten und den eigentlichen bibliographischen Daten. Katalogdaten sind im System des *Online Public Access Catalog* (OPAC) aber nur eine Komponente neben den Daten der Orts- und Fernleihe und den Erwerbungsdaten. Obschon ein Datensatz aus dem Katalog in diesen Bestandteilen zusätzliche Daten erhält, sind diese größtenteils nur innerhalb dieses monolithischen Systems oder ansonsten meist nur über Protokolle wie z.B. Z39.50, das außerhalb der Bibliothekswelt weitgehend unbekannt ist, erreichbar. Auch existiert kein Mechanismus, der es beispielsweise für die Fernleihe ermöglichen würde, Titel anhand ihrer IDs als äquivalent zu kennzeichnen. Selbst die in Web-Browsern für die Benutzer als Ergebnis einer Recherche angezeigten Titeldaten sind als solche nicht Teil des Webs, da sie weder in ihrer Gesamtheit noch in den oben beschriebenen Bestandteilen über URLs für die Indexierung durch Suchmaschinen oder zur Speicherung durch den Benutzer als Lesezeichen verfügbar sind.[1] Des Weiteren bestehen im traditionellen OPAC keine Möglichkeiten, die Titeldaten mit wissenschaftlichen Primärdaten so zu verknüpfen, dass sie in einem gemeinsamen Datenmodell verarbeitet werden könnten. Aber auch administrative Daten über die Bibliothek als Organisation (wie z.B. ihr geographischer Standort oder ihre Öffnungszeiten) liegen zumeist nicht in maschinenlesbarer Form außerhalb des OPAC vor und können nicht zur automatischen Verarbeitung mit den Titeldaten herangezogen werden.

Unter der Prämisse, dass Bibliotheken in der Vergangenheit qualitativ hochwertige strukturierte Daten erzeugt haben, die auch außerhalb dieses Kontextes nachgenutzt werden können, beziehungsweise dass es externe Quellen gibt, deren Daten sowohl für die Bibliotheksbenutzer als auch für das Bibliothekspersonal nutzbringend mit den vorhandenen Informationen verknüpft werden können, hat sich in den letzten Jahren eine internationale Bewegung formiert, die sowohl an den rechtlichen Voraussetzungen für die Freigabe bibliographischer Metadaten arbeitet als auch an der Technologie, mit der unterschiedliche Aspekte der Datenmodellierung umgesetzt werden können. In den folgenden Abschnitten sollen zunächst diese beiden Gesichtspunkte vorgestellt werden, um dann in einigen Anwendungsszenarien zu skizzieren, wie Linked Open Data im Bibliotheksumfeld eingesetzt werden kann.

DIENSTLEISTUNG DER ZUKUNFT: LINKED OPEN DATA

VON **ANDRÉ HAGENBRUCH**

OPEN DATA: RECHTLICHE GRUNDLAGEN

Unter dem Sammelbegriff des „Open"-Paradigmas haben sich in den letzten Jahren Initiativen wie *Open Source, Open Access, Open Content oder Open Government* formiert. Ähnlich wie diese Initiativen fordert die Open Data-Bewegung die uneingeschränkte Veröffentlichung, Nutzbarkeit und Veränderbarkeit von Rohdaten, die von allgemeinem Interesse sein könnten. Ausgehend von der Idee der *Wissensallmende*, dass Informationen anders als natürliche Rohstoffe nicht durch ihre Benutzung aufgebraucht sondern aufgewertet werden[2], versucht diese Bewegung, Daten unterschiedlicher Quellen global im Web verfügbar zu machen, damit diese kollaborativ be- und verarbeit werden können, um beispielsweise in der Wissenschaft den Erkenntnisprozess zu beschleunigen. Organisiert ist die Open Data-Bewegung in Deutschland vor allem im Open Data Network (http://opendata-network.org) und der Open Knowledge Foundation (http://okfn.de), die auch die Plattform Data Hub (http://thedatahub.org) betreibt, auf der zurzeit mehr als 3.000 Datenquellen (davon 52 bibliographische) gehostet werden; nicht alle dieser Quellen stehen jedoch unter einer gemeinfreien Lizenz.

Da bibliographische Metadaten in der Vergangenheit für gewöhnlich nicht explizit mit Lizenzen versehen wurden, galt ihr (urheber)rechtlicher Status als ungeklärt. Mit dem vom *Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen* (hbz) in Auftrag gegebenen rechtlichen Leitfaden *Open Data – Freigabe von Daten aus Bibliothekskatalogen*[3] existiert seit November 2011 eine Orientierungshilfe in der Frage, ob bibliographische Metadaten unter eine gemeinfreie Lizenz gestellt und veröffentlicht werden dürfen. Insbesondere die Frage, wer der Rechteinhaber von Daten ist, die durch Verbundkatalogisierung entstanden sind, ist schwer zu beantworten: Da jede einzelne Bibliothek (zumindest theoretisch) in gleichem Maße in die Erstellung dieser Datenbank investiert hat, müsste jede Einrichtung, die ihren Datenbestand veröffentlichen will, jede andere Verbundbibliothek um Erlaubnis fragen.[4] Die Praxis sieht allerdings anders aus: Nachdem das Hochschulbibliothekszentrum im März 2010 die Daten der Universitäts- und Stadtbibliothek Köln und der Zentralbibliothek der Sporthochschule Köln freigegeben hatte, entschlossen sich auch die Deutsche Zentralbibliothek für Medizin, die Universitätsbibliotheken Aachen, Dortmund und Duisburg-Essen sowie das Landesbibliothekszentrum Rheinland-Pfalz, ihre Kataloge unter der *Creative Commons Public Domain-Lizenz* (CC0) zu veröffentlichen, ohne dass es bei den übrigen Verbundbibliotheken nennenswerten Widerstand dagegen gegeben hätte. Der

gravierendste Einwand gegen eine gemeinfreie Lizenz ist der, dass sie auch eine kommerzielle Verwertung der Daten ermöglicht. Dagegen ist einzuwenden, dass man einerseits oftmals gar nicht entscheiden kann, wann eine kommerzielle Nutzung vorliegt, und dass andererseits eine nicht-kommerzielle Lizenz mit anderen Lizenzen wie z.B. der von der Wikipedia verwendeten Share-Alike-Lizenz aus rechtlichen Gründen nicht kompatibel ist, sodass diese beiden Inhalte nicht miteinander kombiniert werden dürfen.[5] Da aber die Rekombination von Daten und Inhalten mit anderen Informationen der zentrale Punkt des Linked Data-Paradigmas ist, sollte eine Lizenz gewählt werden, die möglichst wenige Restriktionen enthält.

LINKED DATA: TECHNISCHE GRUNDLAGEN

Linked Data wird als Voraussetzung zur Realisierung des *Semantic Web* angesehen, einer Erweiterung des bereits bestehenden Webs. In dieser zukünftigen Version des Web koexistieren Dokumente, die von Menschen gelesen werden können und maschinenlesbare Daten. Linked Data stellt dabei zunächst (eine möglichst große Menge) formal logisch interpretierbarer Aussagen bereit, aus denen dann im Semantic Web durch Inferenzen neue Informationen abgeleitet werden können. Um zu verstehen, weshalb dies einen wichtigen Fortschritt für das Web darstellt, betrachten wir zunächst die aktuelle Situation, um dann über einen Zwischenschritt der syntagmatischen Beschreibung und Strukturierung zur semantischen Beschreibung von Daten zu gelangen.

Dokumente im Web werden in der *Hypertext Markup Language* (HTML) beschrieben, einer Auszeichnungssprache, die hauptsächlich Texte mit Strukturen, wie z.B. Überschriften, Absätzen oder Listen annotiert. Weitere Dokumente können durch Verweise referenziert werden; allerdings sind diese Verweise für gewöhnlich nicht typisiert, d.h. man weiß meistens nicht, um welche Art von Link es sich handelt.[6] Das Web benötigt also einerseits Strukturbeschreibungen generischer Ressourcen, andererseits eine Möglichkeit, allgemeine Relationen zwischen Ressourcen zu beschreiben.

Um das erste Problem zu lösen, wurde die *eXtensible Markup Language* (XML) entwickelt, mit der sich beliebige Objekte durch ein eigenes Vokabular von Elementen und Attributen beschreiben lassen. So kann man beispielsweise in einem bibliographischen Titeldatensatz einen Hauptsachtitel *titel* nennen oder eine Person durch die Elemente *vorname* und *nachname* beschreiben. Die syntagmatische und hierarchische Anordnung der Elemente lässt sich durch eine *Document Type Definition* (DTD) oder ein *XML-Schema* spezifizieren, sodass ein entsprechendes XML-Dokument auch als *valide* bezeichnet wird, wenn es erfolgreich gegen eine solche Spezifikation geparst werden kann. Liegt eine solche Spezifikation nicht vor, kann das Dokument aber dennoch erfolgreich geparst werden, nennt man es *wohlgeformt*. Um Kollisionen zwischen gleichen Benennungen zu vermeiden (z.B. bei der gleichzeitigen Verwendung von *titel* als Hauptsachtitel und für den akademischen Grad einer Person), wurde das Konzept des *Namensraums* eingeführt, mit dem sich unterschiedliche Schemata eindeutig identifizieren lassen. Was diese Mechanismen allerdings nicht leisten, ist eine *semantische* Disambiguierung dieser Benennungen: Ein Sprecher des Deutschen kann durch sein sprachliches Wissen und sein Weltwissen schließen, worum es sich bei diesen

beiden Elementen handeln soll, da aber Bezeichnungen völlig arbiträr vergeben werden können, hätten wir die Elemente auch t1 und p25 nennen können, sodass diese Möglichkeit der Interpretation wegfällt. Einer maschinellen Verarbeitung versperrt sie sich gänzlich, da es sich für den Computer nur um Zeichenketten handelt, die keine weitere Bedeutung besitzen.

Zur Lösung dieses Problems wurde das *Resource Description Framework* (RDF) entwickelt. In ihm stellt jedes zu beschreibende Objekt eine Ressource (auch *Subjekt* genannt) dar, die Eigenschaften (*Prädikate*) besitzt, die wiederum Werte (*Objekte*) haben. Subjekt-Prädikat-Objekt bezeichnet man auch als Tripel; alle Tripel zusammengenommen bilden einen gerichteten Graphen.[7] Für das Subjekt und das Prädikat gilt die Einschränkung, dass sie aus einem *Uniform Resource Identifier* (URI)[8] bestehen müssen, während ein Objekt ein URI oder ein Literal sein kann. Ein Literal kann durch einen optionalen Datentyp (z.B., um anzuzeigen, dass es sich bei der Zeichenkette um ein Datum nach ISO 8601 handelt) oder eine wiederum optionale Sprachangabe nach RFC 3066 näher klassifiziert werden. Darüber hinaus verfügt RDF genauso wie XML über den Mechanismus des Namensraums, aus dem sich unterschiedliche Vokabulare für die Beschreibung einer Ressource verwenden lassen. Wir können nun beispielsweise über den Titel *Allgemeine und molekulare Botanik* folgende Aussagen treffen (siehe Turtle-Notation Abb. unten).

Zunächst deklarieren wir den Namensraum der *Dublin Core Terms*, den wir für unsere Prädikate einsetzen. Während wir im ersten Tripel im Objekt ein Literal verwenden, ist der Wert des Verfassers im zweiten Tripel eine Ressource. Diese lässt sich nun anhand der *SPARQL Protocol and RDF Query Language* (SPARQL), einer SQL-ähnlichen Abfragesprache für Tripel, von der Deutschen Nationalbibliothek holen und mit ihren Tripeln weiter verarbeiten.[9]

Woher weiß man aber nun beispielsweise, ob es sich beim Objekt um ein Literal oder eine Ressource handeln wird? Dazu gibt es die Möglichkeit, Vokabulare mittels *RDF Schema* (RDFS) und / oder der *Web Ontology Language* (OWL) semantisch zu beschreiben. Während sich RDFS eher auf die Definition grundlegender Eigenschaften beschränkt, basiert OWL auf der *Beschreibungslogik*, einer Untermenge der *Prädikatenlogik erster Stufe* zur Wissensrepräsentation. In RDFS lässt sich beispielsweise spezifizieren, dass ein Verfasser eine Ressource sein muss (<http://purl.org/dc/terms/creator> rdfs:range <http://purl.org/dc/terms/Agent>) oder dass diese Eigenschaft eine Spezialisierung der Relation *Mitarbeiter* (<http://purl.org/dc/terms/creator> rdfs:subProperty <http://purl.org/dc/terms/contributor>) ist. In OWL lässt sich darüber hinaus z.B. definieren, dass Verfasser aus dem *Dublin Core*-Vokabular synonym zu Verfassern aus dem *Friend of a Friend* (FOAF)-Vokabular (<http://purl.org/dc/terms/creator> owl:equivalentProperty <http://xmlns.com/foaf/0.1/maker>) sind, dass unsere lokale Ressource äquivalent zu einer externen Ressource ist (<http://data.ub.rub.de/resource/HT015416217> owl:sameAs <http://openlibrary.org/works/OL16322537W/>) oder dass ein Identifier wie z.B. eine ISBN eine owl:InverseFunctionalProperty sein muss, um sie als eineindeutig zu kennzeichnen. Diese Funktionalitäten erhöhen die Interoperabilität unterschiedlicher Datenmodelle, indem sie ein maschinenlesbares, formal verifizierbares Mapping zwischen den Vokabularen ermöglichen.

Ein weiteres wichtiges Merkmal des RDF-Modells ist die *Open World Assumption* (OWA): Obschon es für die Ressource mit dem Titel *Allgemeine und molekulare*

```
@prefix dcterms: <http://purl.org/dc/terms/>.
<http://data.ub.rub.de/resource/HT015416217>  dcterms:title "Allgemeine und molekulare Botanik"@de;
                                              dcterms:creator <http://d-nb.info/gnd/124104177>.
```

Abb.: Turtle-Notation

Botanik drei Verfasser gibt, ist es nicht falsch, dass wir im obigen Beispiel nur einen von ihnen genannt haben. Informationen über weitere Autoren sind lediglich noch nicht bekannt, d.h., es kann nicht geschlossen werden, dass ein Autor, nur weil er hier nicht genannt wurde, nicht Verfasser dieses Werks ist. Unter einer *Closed World Assumption* hingegen müsste man davon ausgehen, dass dieses Beispiel eine vollständige Beschreibung der Ressource ist, sodass eine Anfrage nach einem der anderen Autoren dieser Ressource nicht die leere Menge als Antwort hätte, sondern den Wahrheitswert falsch.

Neben der Möglichkeit, semantische Daten über einen SPARQL-Endpoint im Web zur Verfügung zu stellen, existieren mit *Microdata* und *RDFa* zwei Ansätze, um diese in HTML einzubetten.[10] Während sich der erste auf das aktuell in Entwicklung befindliche HTML5 beschränkt, ist RDFa sowohl für diese Version als auch für XHTML und HTML 4.01 geeignet. Beide Ansätze sind zu RDF kompatibel, jedoch bemüht sich Microdata eher darum, für den Benutzer einfach verständlich zu sein[11]: im Sommer 2011 haben sich die großen Web-Suchdienste Google, Yahoo! und Microsoft (Bing) zur Plattform http://schema.org formiert, auf der sie Schemata für die Verwendung mit Microdata veröffentlichen, die von ihren Suchmaschinen erkannt und verarbeitet werden. Schon jetzt finden sich dort beispielsweise mit Book, ScholarlyArticle, Person oder Library Klassen, die für semantische Beschreibungen im Bibliothekskontext geeignet sind. Mit http://schema.rdfs.org existiert ein Dienst, der einerseits Mappings von bereits vorhandenen Vokabularen wie z.B. Dublin Core, FOAF, GoodRelations oder Bibliographic Ontology auf Schema.org-Vokabulare bereit hält, andererseits täglich aktualisierte Transformationen der Schema.org-Vokabulare auf RDFS/OWL erzeugt. Welche der beiden Varianten zu präferieren ist, lässt sich zu diesem Zeitpunkt nicht eindeutig klären. Wichtig ist aber, dass beide Ansätze die Extraktion strukturierter Informationen auf eine stabile Weise ermöglichen: Während man beim traditionellen Prozess des *Screen Scraping* von Webseiten darauf angewiesen ist, dass sich die Dokumentenstruktur im Laufe der Zeit nicht ändert, braucht man für das Parsing von Microdata bzw. RDFa jeweils nur einen generischen Parser, der auf alle entsprechend ausgezeichneten HTML-Dokumente angewandt werden kann.

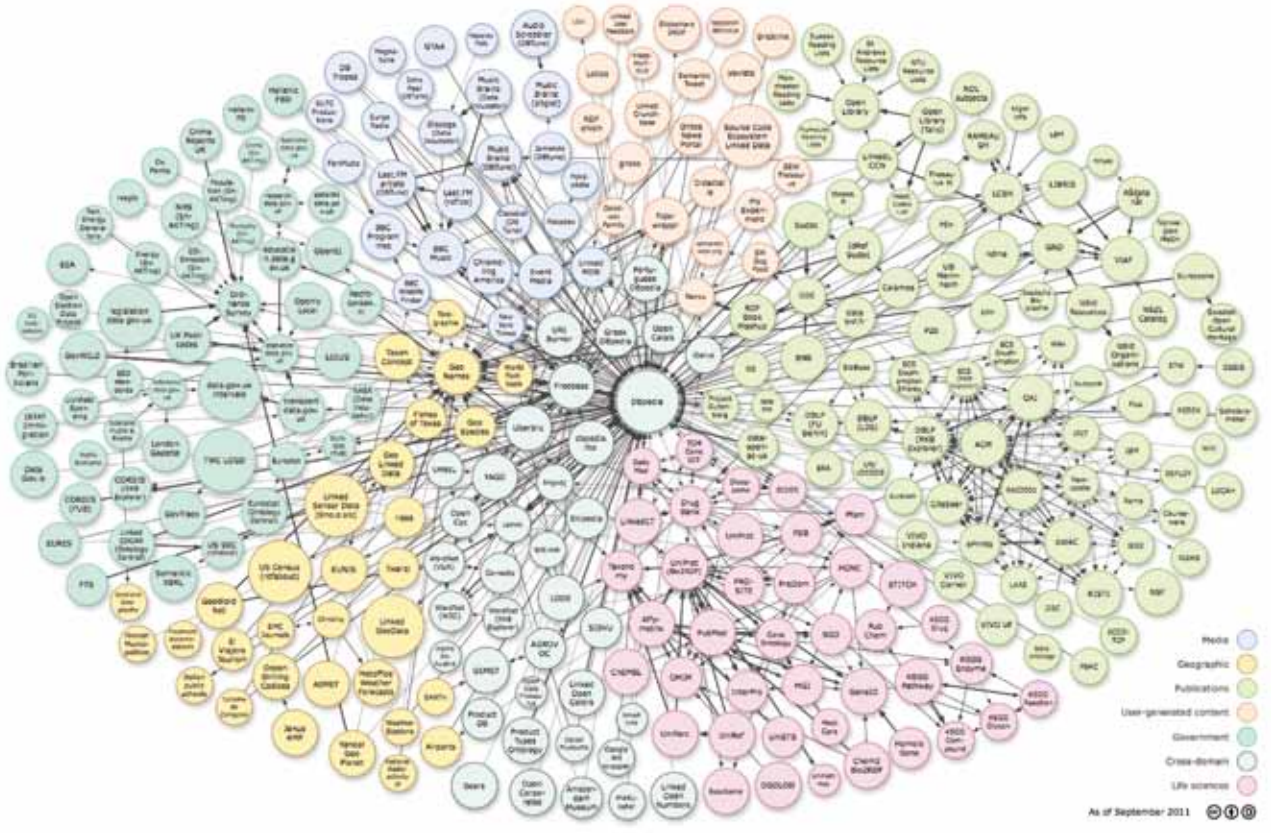
ANWENDUNGSSZENARIEN

Wie wir bereits im letzten Abschnitt gesehen haben, lassen sich unterschiedliche Metadaten auf verschiedene Weisen sinnvoll miteinander verknüpfen. Welche Verknüpfungen zwischen bereits im *Web of Data* veröffentlichten Daten bestehen, lässt sich am seit 2007 regelmäßig von Richard Cyganiak und Anja Jentzsch unter http://lod-cloud.net/ veröffentlichten *Linking Open Data Cloud Diagram* (Abbildung) ersehen:

Zentraler Anknüpfungspunkt ist die *DBpedia* (http://dbpedia.org/), ein Gemeinschaftsprojekt der FU Berlin, der Universität Leipzig und der Firma OpenLink Software, in dem in regelmäßigen Abständen[12] aus Dumps der Wikipedia strukturierte Daten aus den Infoboxen extrahiert und in RDF konvertiert werden.[13] Zum gegewärtigen Zeitpunkt sind in ihr folgende Daten enthalten:

> The new DBpedia data set describes more than 3.64 million things, of which 1.83 million are classified in a consistent ontology, including 416,000 persons, 526,000 places, 106,000 music albums, 60,000 films, 17,500 video games, 169,000 organizations, 183,000 species and 5,400 diseases.
>
> The DBpedia data set features labels and abstracts for 3.64 million things in up to 97 different languages; 2,724,000 links to images and 6,300,000 links to external web pages; 6,200,000 external links into other RDF datasets, and 740,000 Wikipedia categories. The dataset consists of 1 billion pieces of information (RDF triples) out of which 385 million were extracted from the English edition of Wikipedia and roughly 665 million were extracted from other language editions and links to external datasets.[14]

Über die dort enthaltenen bibliographischen IDs (z.B. OCLC-Nr. oder ISBN) oder Personen-IDs lassen sich zusätzliche (auch mehrsprachige) Informationen für die Recherche und die Anzeige in Bibliothekskatalogen



nutzen. Beispielsweise verwendet das *SLUBsemantics*-Projekt (http://www.slub-dresden.de/SLUBsemantics) der TU Dresden einerseits sowohl Übersetzungen der Terme ihres Suchmaschinenindex als auch semantische Relationen wie z.B. die der Synonymie, um die Trefferquote in ihrer Breite (Recall) bei Anfragen an ihren Katalog zu verbessern, andererseits lässt sich die Qualität der Treffer (Precision) erhöhen, indem Mehrdeutigkeiten von Begriffen anhand der Anreicherung der Dokumente mit semantischen Informationen zur Disambiguierung (z.B. *Python* als Programmiersprache, als Schlangenart oder als Kurzbezeichnung der englischen Comedy-Gruppe Monty Python) durch den Benutzer mittels nachgelagerter Filter (sogenannter Facetten) aufgelöst werden können.

Darüber hinaus ließe sich durch die Integration administrativer Daten, z.B. über Öffnungszeiten und geographische Lage einer Bibliothek, das Ranking der Treffer beeinflussen, wenn der physische Zugang zu den Medien relevant ist: In einem Szenario, in dem ein Benutzer eine Recherche beispielsweise über alle UAMR-Bibliotheken ausführt, könnten aktuelle Zeit- und Ortsinformationen des Benutzers Einfluss auf das Ranking haben. Dass die Formalisierung solcher Angaben teilweise nicht trivial ist, sieht man in der folgenden Abbildung:

Während wir als Menschen kein Problem haben, die geographische Lage auf dem Campus anhand von Karten zu interpretieren, lassen sich für die maschinelle Verarbeitung für das Gebäude zwar Geo-Koordinaten angeben, wie sie z.B. für die UB Bochum unter http://lobid.org/organisation/data/DE-294.rdf unter dem Prädikat geo:location zu finden sind, die mit aktuell vorhandenen Geräten über einen Dienst wie GPS verwertbar wären, die genaueren Angaben für die Eingänge (d.h., die Etagen und die Himmelsrichtung) aber sind für eine automatische maschinelle Verarbeitung nicht so gut geeignet. Ähnliches gilt für die Zeiträume *Semester* und *Vorlesungsfreie Zeit* und die Feiertage: während es uns Menschen relativ leicht fällt, diese beweglichen Zeiträume oder Zeitpunkte nachzuschlagen, bedarf es konkreter Datums- und Zeitangaben, die – im Falle der Information über die Semesterzeiträume z.B. von einem SPARQL-Endpoint der Universitätsverwaltung bezogen werden könnten – die sich beispielsweise anhand der *openingHours*-Spezifikation der *GoodRelations*-Ontologie angeben ließen.
(http://www.heppnetz.de/ontologies/goodrelations/v1.html#OpeningHoursSpecification)

Kombiniert man diese Angaben mit den Ausleihinformationen, die sich im RDF-Format der *Document Availability Information API* (DAIA) (http://www.gbv.de/wikis/cls/DAIA_-_Document_Availability_Information_API) transportieren lassen, könnte man in einer gemeinsamen UAMR-Trefferliste eine Sortierung realisieren, welche die optimale Erreichbarkeit eines Mediums anhand der Parameter Datum / Uhrzeit, Standort des Benutzers und Ausleihstatus berücksichtigt. Auch lässt sich das DAIA-Datenmodell für die Orts- und Fernleihe als Abstraktionsschicht über beliebige Lokalsysteme einsetzen, um unabhängig von der jeweiligen Anwendung Informationen darüber liefern zu können, in welchem Status sich ein Medium befindet. Dieser Dienst wird bereits von einigen Bibliotheken im Bibliotheksverbund GBV eingesetzt und wird von einigen Anbietern kommerzieller Suchmaschinenindizes (wie z.B. *Summon* von Serials Solutions oder *Ebsco Discovery Service*) unterstützt.

In der Erwerbung ließen sich durch den Einsatz von Linked Open Data ebenfalls Verbesserungen sowohl für Bibliotheksbenutzer als auch -mitarbeiter erzielen. Seit Mitte Oktober 2011 bietet das Hochschulbibliothekszentrum auf seiner Plattform lobid.org neben Organisations- und Titeldaten auch URIs zu allen Einträgen der Zeitschriftendatenbank (ZDB) der Deutschen Nationalbibliothek an.[15] Auf dieser Grundlage ließe sich z.B. unter Zuhilfenahme der Ergebnisse des Projekts *Shared ERM Requirements* (http://sconulerm.jiscinvolve.org/wp/) mit der eigenen Erwerbungsabteilung ein *Electronic Resource Management* (ERM)-System auf Basis eines populären Web-Frameworks (wie z.B. Ruby On Rails oder Django) entwickeln, das einerseits die bisherige Praxis der Speicherung in Excel-Tabellen ablösen könnte, andererseits wesentlich kostengünstiger als eine kommerzielle Lösung wäre.

Gleichzeitig haben sich in der Erwerbung von Print-Medien in den letzten Jahren sogenannte *Warenkorbsysteme* etabliert. Leider sind auch diese monolithische Deep Web-Anwendungen, die es den Fachreferenten beispielsweise nicht ermöglichen, Titel verschiedener Anbieter automatisch zu aggregieren, um bei Titeln, die nicht der Preisbindung unterliegen, das beste Angebot (möglicherweise sogar automatisch) auszuwählen. Auch ist es nicht möglich, die Fachhierarchien, nach denen die monatlichen Vorschlagslisten erstellt werden, zu aggregieren, um den Fachreferenten auf einen Blick eine vollständige Übersicht aller Neuerscheinungen zu verschaffen. Während für das erste Problem wiederum die GoodRelations-Ontologie Lösungsansätze böte, ließen sich die unterschiedlichen Fachhierarchien im *Simple Knowledge Organization System* (SKOS) (http://www.w3.org/2004/02/skos/) abbilden und mit den im letzten Abschnitt beschriebenen Mechanismen aufeinander oder auf eine allgemeine Klassifikation abbilden.

FAZIT

Mit der Bibliographic Framework Transition Initiative hat die Library of Congress im Frühjahr 2011 einen Paradigmenwechsel hinsichtlich der Verarbeitung bibliographischer Metadaten eingeleitet. Hauptaugenmerk wird dabei auf eine Datenpraxis gelegt, die nicht mehr vom Diktum eines vollständig beschriebenen und abgeschlossenen bibliographischen Datensatzes ausgeht, sondern von einer offenen Menge von Aussagen über Ressourcen (und die schließt auch nicht-bibliographische Fakten ein), die in unterschiedlichsten Kontexten (v.a. aber im Web) verarbeitbar sind. Die UB Bochum hat ihre Katalogdaten im Februar 2012 unter einer gemeinfreien Lizenz veröffentlicht, sodass diese nun im Rahmen der LOD-Initiative des hbz als Linked Open Data verfügbar sind. Damit ist ein Grundstein gelegt, um diese Daten in neuen Kontexten nutzbar zu machen. Mittel- und langfristig verfolgen wir eine Strategie, Linked Open Data in möglichst viele Projekte zu integrieren und werden auch bei der Auswahl neuer Softwaresysteme auf ihre Kompatibilität zu diesem Paradigma achten.

André Hagenbruch ist Software-Entwickler und Projektmanager in der Universitätsbibliothek Bochum.

ENDNOTES

[1] Vgl. CHRISTIAN HAUSCHKE, Permalinks für Katalogisate. Blog, 2009. http://infobib.de/blog/2009/10/27/permalinks-fur-katalogisate/ (abgerufen am 05.01.2012).
[2] Vgl. DAVID BOLLIER, The growth of the commons paradigm, in: Understanding knowledge as a commons. from theory to practice. Workshop on Scholarly Communication as a Commons ; (Bloomington, Indiana Univ.) : 2004.03.31-04-02, hg. v. CHARLOTTE HESS, ELINOR OSTROm, Cambridge, Mass 2007, S. 27–40.
[3] Vgl. TILL KREUTZER, Open Data – Freigabe von Daten aus Bibliothekskatalogen, Köln 2011.
[4] Ebd., S.29-30
[5] ADRIAN POHL, Open Data im hbz-Verbund, in: ProLibris, 15, H. 3 2010, S. 109–113, S. 110.
[6] Mit den Attributen rel und rev für Links steht ein begrenztes Vokabular zur Beschreibung von Verweisen zur Verfügung (vgl. http://de.selfhtml.org/html/verweise/typisierte.htm). Da es sich hierbei aber nicht um generische Relationen handelt, ist dieses Vokabular nur sehr eingeschränkt für die semantische Annotation einzusetzen.
[7] Dies bedeutet, dass die Eigenschaften einer Ressource nicht geordnet sind, und dass es unterschiedliche Serialisierungen eines solchen Graphen geben kann. Die kanonische Form ist die Darstellung der Tripel in RDF/XML, da diese aber sehr umfangreich und für Menschen schwierig zu lesen sein kann, verwenden wir für unsere Beispiele die wesentlich kompaktere *Turtle*-Notation.
[8] Ein URI stellt eine Abstraktion über eine URL dar: Er identifiziert eine Ressource eindeutig, muss aber nicht unbedingt über ein Protokoll wie HTTP referenzierbar sein.
[9] Gespeichert werden diese Ressourcen in sogenannten *Triple Stores*, dies sind für dieses Datenmodell optimierte Datenbanken. Angegliedert ist ihnen meist ein sogenannter *SPARQL-Endpoint*, über den sich derlei Anfragen absetzen lassen. Leider verfügt die DNB zum jetzigen Zeitpunkt noch nicht über einen öffentlichen SPARQL-Endpoint, doch könnte man den von ihr zur Verfügung gestellten Daten-Dump der *Gemeinsamen Normdaten-Datei* (GND) in einen lokalen Tripel Store importieren, um aus diesem die Informationen zu erhalten.
[10] Mit *Microformats* (http://www.microformats.org) existiert ein weiterer Mechanismus, um Elemente mittels der standardmäßig vorhandenen HTML-Attribute class und rel mit rudimentären semantischen Informationen zu annotieren. Rudimentär ist dieser Ansatz insofern, als er einerseits auf eine begrenzte Anzahl von Vokabularen festgelegt ist, die durch einen gemeinschaftlichen Entscheidungsprozess entwickelt werden, und es andererseits keine uniforme Möglichkeit des Parsings und der formalen Validierung gibt. Zwar gibt es in jüngster Zeit mit *microformats 2.0* (http://www.microfomats.org/wiki/microformats-2) einen ersten Entwurf, um diese Probleme zu beseitigen, da es aber für bibliographische Metadaten ebenfalls nur einen Entwurf gibt, räumen wir diesem Ansatz hier nur geringen Raum ein.
[11] Mit *RDFa Lite*, einer Untermenge von RDFa 1.1, existiert seit November 2011 eine Variante, die den Einstieg in semantisches Markup erleichtern soll.
[12] Seit dem Sommer 2011 existiert mit *DBpedia Live* (http://live.dbpedia.org/) auch die Möglichkeit, unabhängig vom Turnus der Datendumps kontinuierlich aktuelle Daten aus der Wikipedia transformieren und nachnutzen zu können
[13] Anfang 2012 startet die Wikimedia Deutschland ein Projekt mit dem Titel *WikiData* (http://meta.wikimedia.org/wiki/New_Wikidata), das den einzelsprachigen Wikipedien eine zentrale Faktendatenbank zur Verfügung stellen soll, mit der Infoboxen automatisch befüllt werden können. Eines der Exportformate soll dabei RDF sein.
[14] CHRIS BIZER, DBpedia 3.7 released, including 15 localized editions. Blog, 2011. http://blog.dbpedia.org/2011/09/11/dbpedia-37-released-including-15-localized-editions/ (abgerufen am 02.01.2012).
[15] Von den derzeit knapp 1,6 Millionen Datensätzen liegen zu ca. 300.000 detaillierte Beschreibungen vor; der Rest liegt nicht als Open Data vor.

LITERATURVERZEICHNIS

BIZER, CHRIS, DBpedia 3.7 released, including 15 localized editions. Blog, 2011. http://blog.dbpedia.org/2011/09/11/dbpedia-37-released-including-15-localized-editions/ (abgerufen am 02.01.2012).
BOLLIER, DAVID, The growth of the commons paradigm, in: Understanding knowledge as a commons. from theory to practice. Workshop on Scholarly Communication as a Commons ; (Bloomington, Indiana Univ.) : 2004.03.31-04-02, hg. v. Charlotte Hess, Elinor Ostrom, Cambridge, Mass 2007, S. 27–40.
HAUSCHKE, CHRISTIAN, Permalinks für Katalogisate. Blog, 2009. http://infobib.de/blog/2009/10/27/permalinks-fur-katalogisate/ (abgerufen am 05.01.2012).
KREUTZER, TILL, Open Data – Freigabe von Daten aus Bibliothekskatalogen, Köln 2011.
LOUGEE, WENDY PRADT, Scholarly communication and libraries unbound:. the opportunity of the commons, in: Understanding knowledge as a commons. from theory to practice. Workshop on Scholarly Communication as a Commons ; (Bloomington, Indiana Univ.) : 2004.03.31-04-02, hg. v. Charlotte Hess, Elinor Ostrom, Cambridge, Mass 2007, S. 311–332.
NIELSEN, MICHAEL A., Reinventing discovery. the new era of networked science, Princeton 2011.
POHL, ADRIAN, Open Data im hbz-Verbund, in: ProLibris, 15, H. 3 (2010), S. 109–113.