Zur Konzeption und Implementierung einer Infrastruktur für freie bibliographische Daten

Adrian Pohl & Felix Ostrowski (Hochschulbibliothekszentrum Nordrhein-Westfalen)

Abstract

Das zunehmende Bewusstsein für „Open Data“ in der Bibliothekswelt eröffnet wichtige Fragen bezüglich des Umgangs mit freien Daten. Der vorliegende Text diskutiert die konzeptionellen Grundlinien einer technischen Open-Data-Infrastruktur und arbeitet die vielschichtigen Anforderungen an eine solche Infrastruktur heraus. Die behandelten Gesichtspunkte reichen von der Datenpublikation über ihre Beschreibung bis hin zur Änderungsverwaltung. Für einige Aspekte werden vielversprechende Anknüpfungspunkte identifiziert, z.B. in Gestalt von Versionsverwaltungstools aus der Open-Source-Community oder in Form von laufenden Projekten der Open Knowledge Foundation oder des W3C.

Vorwort

Mit dem Internet und - als dessen wichtigstem Bestandteil - dem World Wide Web formt sich seit einigen Jahrzehnten eine umfassende Publikations- und Kommunikationsplattform aus, auf der zukünftig der Großteil aller Publikation und Kommunikation stattfinden wird. Als eine Erweiterung des bestehenden Webs lässt sich Linked Open Data verstehen. Mit Linked Open Data werden zwei Standards bezeichnet, die die Funktionalität eines Netzes von Daten sichern sollen, indem sie die rechtliche und technische Kompatibilität von Daten im Web garantieren:

- Open-Data-Standards sorgen für die *rechtliche Basis* der Nutzung und Kombination verteilter Daten im Netz.
- Linked-Data-Standards sorgen für die *technische* Kompatibilität zwischen verteilt vorliegenden Daten.

In einer dreiteiligen Artikelreihe über Linked-Open-Data-Aktivitäten am Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz) sollen die rechtlichen wie technischen Dimensionen von Linked Open Data erläutert werden und die Notwendigkeit, die Ziele und der Nutzen von Linked Open Bibliographic Data1 dargelegt werden. Im ersten Teil dieser Reihe über das Was, Warum und Wie von Linked-Open-Data-Aktivitäten am hbz sollen einige Fragen zu Open Data geklärt werden. Er erscheint gedruckt in ProLibris 3/2010. Der zweite Teil – gemeinsam verfasst von Felix Ostrowski und Adrian Pohl – mit dem Schwerpunkt Linked Data erscheint gedruckt in B.I.T. online 3/2010 und der dritte, in dem sich ebenfalls Felix Ostrowski und Adrian Pohl mit der Konzeption und Implementierung einer Open-Data-Infrastruktur befassen, wird gedruckt im Tagungsband der DGI-Konferenz *Semantic Web & Linked Data Elemente zukünftiger Informationsinfrastrukturen* publiziert. Alle Texte werden darüber hinaus unter einer CC-BY-Lizenz im Web publiziert, siehe etwa unter http://www.hbz-nrw.de/dokumentencenter/produkte/lod/.

1 Einleitung

Anfang 2010 haben mehrere Bibliotheken ihre Katalogdaten vollständig der Öffentlichkeit zur freien Verfügung gestellt: die CERN Library im Januar1, die Universitätsbibliothek Ghent im Februar2 und im März schließlich mehrere Kölner Bibliotheken und das Landesbibliothekszentrum Rheinland-Pfalz in Kooperation mit dem Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen3. So begrüßenswert diese Aktivitäten sind, klar ist: Dies ist erst der Anfang, denn mit der Publikation von Rohdaten allein ist es nicht getan. Die Veröffentlichung der Daten unter einer freien Lizenz ist sicher ein notwendiger politisch-rechtlicher Schritt, der wichtige Signale setzt. Neben den rechtlichen und politischen Aspekten von Open Data ist es aber ebenso wichtig, zum einen die Entstehung einer Open-Data-Community und -Praxis zu fördern und zum anderen eine technische Open-Data-Infrastruktur zu entwickeln, welche diese Open-Data-Praxis unterstützt. Denn sobald eine gewisse Anzahl an Institutionen ihre Daten freigibt, sollte eine entsprechende Infrastruktur dabei helfen, den Nutzen freier Katalogdaten zu maximieren, um eine tragfähige und zukunftsträchtige Open-Data-Praxis in der Bibliothekswelt zu entwickeln.

Im hbz wird im engen Austausch mit Beteiligten und Interessierten an der Konzeption und Implementierung einer Open-Data-Infrastruktur gearbeitet, welche die offenen Fragen der Beschreibung, Aktualisierung und Versionierung von Rohdaten adressiert. Dabei wird an die Erfahrungen, Entwicklungen und Praktiken aus der Open-Source-Community wie aus anderen relevanten Projekten angeknüpft.[4]

Es handelt sich hier in erster Linie um Überlegungen für eine generische Open-Data-Infrastruktur, die nicht allein für Linked Data[5] konzipiert wird. Aber sie kann eben auch einer Linked-Open-Data-Praxis dienen. Dies könnte in der Übergangsphase zum Semantic Web eine wichtige Rolle spielen.

Da zu dieser Thematik bisher sehr wenig Texte zu finden sind, wird hier erst einmal tentativ der Rahmen der ganzen Fragestellung abgesteckt. In diesem Text werden zunächst die grundlegenden Fragen herausgearbeitet, die bei der Konzeption einer solchen Infrastruktur beantwortet werden müssen. Im zweiten Schritt werden relevante bestehende Projekte erläutert, die zur Implementierung einer solchen Infrastruktur nachgenutzt werden könnten. Im letzten Schritt stellen wie ein erstes Konzept einer solchen Infrastruktur vor.

---

1 Siehe die Pressemtitteilung CERN (2010): The CERN Library publishes its book catalog as Open Data. Einsehbar unter http://library.web.cern.ch/library/Library/announcement.html [05.08.2010].

2 http://lib.ugent.be/info/en/exports.shtml [05.08.2010]

3 Siehe die Pressemitteilung Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (2010): Freigabe der Katalogdaten: Kölner Bibliotheken leisten Pionierarbeit. Pressemitteilung, einsehbar unter: http://www.hbz-nrw.de/dokumentencenter/presse/pm/datenfreigabe [05.08.2010] und Pohl, Adrian (2010): Open Data im hbz-Verbund. Erscheint in ProLibris 3/2010. Preprint einsehbar unter http://www.hbz-nrw.de/dokumentencenter/produkte/lod/ [05.08.2010].

4 Im März veröffentlichte Nat Torkington im O'Reilly Radar einen Beitrag mit dem Titel „Truly Open Data", der die Notwendigkeit konstatiert, für Open Data etablierte Verfahren der Open-Source-Gemeinschaft zu übernehmen, seien dies der kollaborative Ansatz und die Rolle von Kuratoren (Maintainern) oder Dokumentation, Bug Tracking und Versionierung. Dieser Text trug maßgeblich zur Beschäftigung mit dem Thema im hbz bei. Siehe Torkington, Nat (2010): Truly Open Data. Einsehbar unter http://radar.oreilly.com/2010/03/truly-open-data.html [05.08.2010].

5 Vgl. Pohl, Adrian / Ostrowski, Felix (2010): "Linked Data" - und warum wir uns im hbz-Verbund damit beschäftigen. Erscheint in ProLibris 3/2010. Preprint einsehbar u.a. unter http://www.hbz-nrw.de/dokumentencenter/produkte/lod/ [05.08.2010].

2 Fragen zur Konzeption

In diesem Abschnitt werden grundlegende Fragen zu Zweck, Funktionalitäten, Standards und Architektur einer Open-Data-Infrastruktur ausgesprochen, die vor jeder Implementierung beantwortet werden sollten.

2.1 Zweck

Welchen Zwecken soll eine Open-Data-Infrastruktur genügen, welche Aufgaben erledigen? Welche Aktivitäten soll sie unterstützen? Allgemein können hier drei Einsatzbereiche einer Open-Data-Infrastruktur unterschieden werden: *Publikation*, *Nachnutzbarkeit* und *Dokumentation*.

2.1.1 Publikation

Minimale Voraussetzung, um etwas als Open Data bezeichnen zu können, ist die Veröffentlichung von Daten unter einer offenen Lizenz[6] im Internet. Eine Open-Data-Infrastruktur sollte es also jeder Institution (und jeder Privatperson) ermöglichen, ihre Daten online unter einer Open-Data-Lizenz zu publizieren, auch jenen, für die der Betrieb einer entsprechenden Plattform auf einem eigenen Webserver zu aufwändig ist.

2.1.2 Nachnutzung

Wie bereits erwähnt, ist eine Datenpublikation nur der notwendige erste Schritt. Damit die Daten auch nachgenutzt werden können, müssen sie von Interessierten gefunden und auf einfache Art heruntergeladen werden können.

2.1.3 Dokumentation

Um die Vertrauenswürdigkeit von Daten zu sichern, aber auch, um den Grad der Nachnutzung bestimmter Daten nachvollziehen zu können, ist es wichtig, die Provenienz von Datensammlungen zu dokumentieren. Im Idealfall lässt sich zu jeder Datensammlung jederzeit nachweisen, wo die Daten ihren Ursprung haben und wer wann an ihnen etwas geändert oder ergänzt hat. Auch sollte berücksichtigt werden, dass Publizierende häufig nachvollziehen wollen, wer ihre Daten in welchem Kontext nachnutzt. Diese Anforderungen lassen sich nicht allein technisch erfüllen, sondern sind auch und gerade eine Frage von sozialen Konventionen.

2.2 Funktionen

Welche Funktionalitäten müssen implementiert werden, um die im vorherigen Abschnitt genannten Zwecke einer Open-Data-Infrastruktur zu erfüllen? Welches sind notwendige Basisfunktionalitäten und welches wünschenswerte weitere Funktionalitäten?
Unter diesen Gesichtspunkten werden im Folgenden die ermittelten Grundfunktionalitäten - *Speicherung/Publikation*, *Beschreibung*, *Aggregierung*, *Verlinkung*, *Aktualisierung*, *Versionierung* und *Download-Tracking* - näher betrachtet.

6 Zur Open-Data-Lizenzierung vgl. Pohl, Adrian (2010), Abschnitt 4

2.2.1 Speicherung/Publikation

Eine Open-Data-Infrastruktur sollte es allen - auch Institutionen, die keinen eigenen Webserver zu ihrer Verfügung haben - ermöglichen, Datensammlungen in unterschiedlichen Formaten unter einer offenen Lizenz zu publizieren.[7] Es müsste demnach ein Dienst geschaffen werden, der es Bibliotheken ermöglicht ihre offenen Datensammlungen hochzuladen.[8]

2.2.2 Beschreibung & Verlinkung

Um die Nachnutzung von Open Data zur ermöglichen, müssen relevante Daten aufgefunden werden können. Der erste Schritt auf dem Weg, Auffindbarkeit herzustellen, ist die Beschreibung relevanter Eigenschaften der Daten. Dazu gehört zuallererst die Dokumentation des verwendeten Datenmodells und -formats. Darüber hinaus sind Auskünfte zur konkreten Datensammlung wichtig. Beispielsweise sind Angaben zum Publikationsdatum und der veröffentlichenden Institution interessant sowie zum Schwerpunkt der Sammlung, die durch die Katalogdaten beschrieben wird. Wichtiger Bestandteil der Beschreibungen ist die Angabe von Beziehungen zwischen Datensammlungen. So verweisen Titeldaten etwa auf Personendaten und andere Normdaten, Systematiken, Exemplardaten verweisen auf Titeldaten usw. Diese Angaben sind für eine effektive Nachnutzung u.U. sehr wichtig.

2.2.3 Aggregierung

Beschreibung allein ermöglicht nicht notwendigerweise das Auffinden der Daten. Beschreibungen von Datensammlungen müssen aggregiert und recherchierbar gemacht werden, um eine optimale Auffindbarkeit zu unterstützen. Es gibt verschiedene Möglichkeiten, dies zu erreichen, etwa durch die Schaffung einer zentralen Plattform, auf der möglichst viele Datensammlungen direkt verzeichnet werden oder durch eine dezentrale Beschreibung und nachträgliche Aggregierung der Metadaten in einem Dienst.

2.2.4 Aktualisierung

Eine Open-Data-Praxis kann nicht allein darin bestehen, Daten einmal mit einer offenen Lizenz auf einen Server zu legen. Da sich die Daten in Bibliothekskatalogen kontinuierlich wandeln, ist es nötig, sie regelmäßig zu aktualisieren. Während dies vor allem eine Anforderung an Daten veröffentlichende Institutionen ist, muss eine Open-Data-Infrastruktur dies technisch unterstützen und entsprechend konzipiert werden.

2.2.5 Versionierung

Eine Aktualisierung kann die Bereitstellung eines neuen Komplettexports bedeuten. Viel sinnvoller wäre es allerdings, nur die Veränderungen zum letzten Export anzugeben, also "Patches" zu liefern, mit denen Interessierte sich ein Update einspielen können, um die kontinuierlichen Veränderungen nachzuvollziehen. Dies ist vergleichbar mit dem Problem der Versionierung von Softwarecode in einem Repositorium oder von Texten in einem Wiki: Unterschiede zwischen Versionen ("Diffs") werden in Patches dokumentiert und ermöglichen es anderen, diese Aktualisierungen zu

7 Gerade kleineren Bibliotheken ist es oft nicht möglich, die Daten auf einem eigenen Server zu speichern.
8 Dies bedeutet im Vorgriff auf Abschnitt 4.1, dass die Open-Data-Infrastruktur zunächst nicht völlig dezentral sein kann, sondern einen Dienst beinhalten sollte, der von allen benutzt werden kann.

übernehmen. Darüber hinaus sorgt die Versionskontrolle dafür, dass die zeitliche Abfolge der Änderungen sowie deren Urheber dokumentiert werden. Es gilt, diese Praktiken in Zukunft auf Daten zu übertragen oder wie Nat Torkington schreibt: *'The users of data will have to adapt to the idea of versions, like the users of software have.'*[9]

2.2.6 Download-Tracking

Für Institutionen, die ihre Daten freigeben, mag es interessant und nützlich sein, zu wissen, wer diese Daten runterlädt und weiterverwendet. Dementsprechend sollte eine Open-Data-Infrastruktur die Möglichkeit berücksichtigen, zumindest nachzuverfolgen, wie oft und von wem Daten heruntergeladen werden.[10] Aufgrund der dezentralen Natur des Webs und dem offenen Charakter der publizierten Daten ergeben sich hier enge Grenzen. Es besteht lediglich die Möglichkeit, Downloads, die unmittelbar vom entsprechenden Server erfolgen, festzuhalten. Da die Daten frei weitergegeben werden können, kann dies zu einer Verbreitung und Nachnutzung führen, die nicht mehr vollständig nachvollziehbar ist.

2.3 Standards

Jede Implementierung einer Open-Data-Infrastruktur sollte die Frage nach den zugrundeliegenden Standards beantworten, u.a.:

- Welches Metadatenmodell liegt der Beschreibung der Datensammlungen zugrunde? Gibt es bestehende Vokabulare, die nachgenutzt werden können?
- In welchem Format sollen die Beschreibungsdaten vorliegen?
- Gibt es Standards zur Versionierung von Daten oder zur Benachrichtigung über Updates?
- …

2.4 Architektur

Welche grundlegenden Entscheidungen müssen hinsichtlich der Architektur einer Open-Data-Infrastruktur gefällt werden? Es lassen sich in dieser Hinsicht drei Möglichkeiten, um die Beschreibung, Publikation und Versionierung von Open Data zu implementieren, identifizieren:

1. Mittels einer *zentralen* Plattform, auf der die Verwaltung von freien Katalogdaten stattfinden soll.
2. Mittels einer *verteilten* Infrastruktur, in der Beschreibung, Speicherung und Versionierung dezentral stattfinden und durch die Benutzung gemeinsamer Standards das einfache Aggregieren der Inhalte ermöglicht wird.
3. Eine *Mischform* aus den ersten beiden, d.h. es würde ein zentraler Dienst bestehen, der die nötigen Funktionen erfüllt, die dahinterliegende Software würde aber auch einzelnen Projekten zur Verfügung gestellt.[11]

---

9 Torkington (2010)

10 Diese Anforderung wurde im Rahmen einer Session auf dem BibCamp³ in Hannover von einem Teilnehmer ergänzt.

11 Diese Variante besteht also sowohl aus der Entwicklung einer geeigneten Software als auch aus dem Anbieten eines Dienstes auf Basis der Software. Damit ist die Möglichkeit gegeben, den bestehenden Dienst zu nutzen, oder aber sich mit einer eigenen Installationeinen in die Open-Data-Infrastruktur einzuklinken. Vgl. hierzu den in Abschnitt 3.1 skizzierten "Software and a Service"-Gedanken, der den Projekten der OKFN zugrunde liegt.

3 Relevante bestehende Projekte

Diese Ideen sind nichts Neues, vielmehr gibt es bereits Projekte, die das ein oder andere Problem schon gelöst haben bzw. an einer Lösung arbeiten. Zwei dieser Projekte sollen hier näher betrachtet werden, da eine Nachnutzung der Ergebnisse bzw. eine Kooperation sehr fruchtbar sein könnte.

3.1 CKAN

Die Open Knowledge Foundation verfolgt bereits seit einigen Jahren das Ziel, freie Daten im Web nachzuweisen, und hat 2008 das Comprehensive Knowledge Archive Network (CKAN) gestartet. „CKAN" bezeichnet sowohl eine Software[12] als auch einen Service[13]. Im Folgenden werden wir mit „CKAN" auf die Software bezugnehmen und mit „CKAN.net" auf den Dienst. Die CKAN-Software ist ein in der Programmiersprache Python geschriebenes Katalogsystem für Open-Data-Pakete oder andere Wissensressourcen, d.h. sie dient der Sammlung von Metadaten über Datenpakete und andere Ressourcen. CKAN ist Open Source und wird u.a. für den Dienst CKAN.net als auch in Open-Government-Data-Katalogen wie http://data.gov.uk eingesetzt. Das Metadatenmodell für die Beschreibung von Datenpaketen ist recht überschaubar[14], allerdings hat CKAN.net den Vorteil, dass die NutzerInnen beliebige Datenelemente ergänzen können. Sämtliche strukturierten Daten werden wie in einem Wiki versioniert. Es gibt ein Web-Interface zum Hinzufügen und Editieren von Paketinformationen. Für den maschinellen Zugriff ist eine API implementiert. Im CKAN-Verzeichnis, das eben nicht auf bibliographische Daten beschränkt ist, gibt es bereits eine Gruppe für bibliographische Daten, siehe http://ckan.net/group/bibliographic. Dort sind auch die Daten aus dem hbz, der Universitäts- und Stadtbibliothek Köln sowie aus der Zentralbibliothek der Sportwissenschaften der Deutschen Sporthochschule Köln verzeichnet. Sowohl die CKAN-Software als auch der Dienst CKAN.net sind für die Implementierung eines Open-Bibliographic-Data-Verzeichnisses attraktiv und sollten bei der Konzeption berücksichtigt werden. Ein Problem mit CKAN ist, dass es allein als ein Verzeichnis fungiert und weder einen Upload noch eine Versionierung von Daten nicht ermöglicht. Ein weiterer Nachteil von CKAN ist, dass die Daten in einer SQL-Datenbank gespeichert werden und bislang nicht als Linked Data in Form von RDF[15] verfügbar sind. Wie in Abschnitt 4.2 erläutert wird, ist dies eine Anforderung, die wir an eine Open-Data-Plattform stellen. Allerdings arbeitet die OKFN daran, CKAN auf RDF-Daten zu basieren und an sogenannte Triple-Stores anschließen zu können:[16] Kürzlich wurde im OKFN-Blog verkündet, dass mit ORDF (OKFN RDF Library) eine Middleware entwickelt wird, mit der die native Speicherung von Daten in Anwendungen wie CKAN direkt und ausschließlich in Triple-Stores stattfinden kann.[17] Eine (Nach-)Nutzung von CKAN für eine Open-Bibliographic-Data-Plattform könnte also durchaus sinnvoll sein.

---

12 http://knowledgeforge.net/ckan/trac [05.08.2010]

13 http://www.ckan.net [05.08.2010]

14 Es gibt folgende Standardmetadatenelemente: Identifier, Titel, Version, URL, Bemerkungen, Tags, Downloadlinks mit Format, Beschreibung und Prüfsumme, Autor, Kurator, Lizenz sowie Bewertungen.

15 Vgl. Pohl/Ostrowski (2010)

16 Ein erster Versuch war der „Semantic CKAN"-Prototyp unter http://world.ckan.net/ [05.08.2010].

17 Pollock, Rufus (2010a): ORDF - the OKFN RDF Library. Einsehbar unter http://blog.okfn.org/2010/07/02/ordf-the-okfn-rdf-library/ [05.08.2010].

3.2 dcat

dcat ist ein RDF-Vokabular, das der Interoperabilität verschiedener Datenkataloge dienen soll. In erster Linie geht es um ein Standardvokabular in RDF für Kataloge zur Verzeichnung von Open *Government* Data. Diese Kataloge gibt es von Nationalstaaten (z.B. UK[18] und USA[19]), von Bundesstaaten (z.B. Maine[20]), von Kommunen und Städten (z.B. London[21] oder San Francisco[22]). Die in diesen Katalogen verzeichneten Daten sind ziemlich heterogen, gemeinsam ist den Daten, dass sie aus der öffentlichen Verwaltung kommen: Ob Gesundheit, Verteidigung, Energie oder Verkehr, es lassen sich die verschiedensten Daten in den Katalogen finden, von Geodaten und Kartenmaterial[23] über beliebte Vornamen[24] bis hin zu Verbrechensstatistiken[25]. Das Problem ist, dass diese unterschiedlichen Kataloge in verschiedenen Formaten und mit verschiedenen Metadaten vorliegen. Dadurch wird Interoperabilität erschwert, zumal die Katalogisierung häufig inkonsistent und die benutzen Metadatenelemente oft nicht dokumentiert sind. Mit dcat soll nun ein Standardvokabular in RDF-Schema für Open-Government-Data-Kataloge entstehen, um deren Interoperabilität zu optimieren. Die Entwicklung des dcat-Vokabulars wurde 2010 von Richard Cyganiak angestoßen, der eine erste Fassung auf den Seiten des Digital Enterprise Research Institute (DERI) veröffentlichte.[26] Mittlerweile wird dcat vom Data Catalog Vocabulary Project, einer Arbeitsgruppe am W3C, weiterentwickelt.[27] Der Entwicklungsprozess ist derzeit noch nicht abgeschlossen, die Arbeitsgruppe hat noch keine offizielle Version des dcat veröffentlicht.[28]

4 Konzeption & Implementierung

Konzeption und Implementierung einer Infrastruktur für Open Data befinden sich noch in ihren Anfängen. Die große Dynamik in diesem Bereich, die etwa in den laufenden Projekten des W3C und der OKFN ihren Ausdruck findet, macht einen flexiblen und kooperativen Angang an das Projekt notwendig. Im Folgenden stellen wir kurz unsere ersten Umsetzungsversuche und Tests zur Thematik vor.

---

18 http://data.gov.uk/data [05.08.2010]
19 http://www.data.gov/catalog/raw [05.08.2010]
20 http://www.maine.gov/cgi-bin/data/index.pl [05.08.2010]
21 http://data.london.gov.uk/ [05.08.2010]
22 http://www.datasf.org/ [05.08.2010]
23 Vgl. Open Data des Ordnance Survey in Großbritannien (http://www.ordnancesurvey.co.uk/oswebsite/opendata/) oder die TIGER-Karten des United States Census Bureaus (http://www.census.gov/geo/www/tiger/tgrshp2009/tgrshp2009.html [05.08.2010]). Das deutsche Bundesamt für Kartographie und Geodäsie (http://www.bkg.bund.de [05.08.2010]) hingegen verkauft seine „Produkte“ (noch) in einem „Online-Shop“.
24 Siehe etwa die Daten über die beliebtesten schottischen Vornamen 2009: http://data.gov.uk/dataset/babies_first_names_scotland [05.08.2010].
25 Siehe etwa hier http://www.data.gov/raw/1583/ [05.08.2010].
26 Siehe http://vocab.deri.ie/dcat [05.08.2010].
27 http://www.w3.org/egov/wiki/Data_Catalog_Vocabulary [05.08.2010]
28 Gleichwohl findet sich ein Entwurf unter http://www.w3.org/egov/wiki/Data_Catalog_Vocabulary/Vocabulary_Reference [05.08.2010].

4.1 Dezentralität als Grundsatz der Architektur

Wir sind der Überzeugung, dass eine dezentrale Datenhaltung zu den Grundsätzen einer Open-Data-Praxis gerechnet werden muß.[29] Eine zentrale Plattform, auf der allein alle Datenpakete beschrieben und publiziert werden, impliziert immer auch die Abhängigkeit von eben dieser. Aus diesem Grund halten wir es auch für unabdingbar, die Beschreibungen gemäß den Linked-Data-Standards bereitzustellen, da diese Standards im Kern dem Aufbau einer verteilten Datenbank dienen.[30] Als Grundlage für weitere Überlegungen gilt daher, dass

1. Daten dezentral publiziert und beschrieben werden[31] und
2. die Beschreibungen aggregiert werden, um sie zentral sichtbar und durchsuchbar zu machen[32].

Dieses Konzept ist grundsätzlich nicht neu. In der Bibliothekswelt funktionieren institutionelle Repositorien analog dazu: Metadaten zu den dort vorhandenen Publikationen können über OAI-Schnittstellen abgerufen und somit aggregiert werden. Neu hingegen ist der Ansatz, keine zusätzliche Schnittstelle im klassischen Sinne zu schaffen, denn fortan wird gelten: "Your Website is Your API"[33].

4.2 Beschreibung

Wir halten es für sinnvoll und notwendig, die Infrastrukur so zu konzipieren, dass alle Beschreibungen von Datensammlungen in RDF vorgenommen werden, denn Linked-Data-Standards sorgen für optimale Interoperabilität.[34] Auf der weniger granularen Ebene der Verzeichnung von rohen Datenbankabzügen würde also schon Linked Data produziert werden, auch wenn die Titel- und Normdaten noch in opaken Formaten wie MAB oder MARC vorliegen.

Welches Vokabular soll zur Beschreibung benutzt werden? Wie bereits im zweiten Teil dieser Artikel-Reihe erläutert, ist es äußerst sinnvoll, bestehende Ontologien für eigene Zwecke nachzunutzen.[35] Im Zusammenhang mit der Beschreibung von offenen Rohdaten wird derzeit - wie in Abschnitt 3.2 erläutert - dcat entwickelt. Wir halten es für sinnvoll, den Nutzen dieses Vokabulars für unsere Zwecke zu evaluieren.[36] Wenn auch Government Data sich von Daten aus Bibliotheks- und Verbundkatalogen - etwa durch eine größere Heterogenität zwische den verschiedenen Datensammlungen - unterscheiden mögen, gibt es dennoch eine Menge Gemeinsamkeiten zwischen Open Government

---

29 Vgl. Pollock, Rufus (2006): The Four Principles of (Open) Knowledge Development. Einsehbar unter http://blog.okfn.org/2006/05/09/the-four-principles-of-open-knowledge-development/ [05.08.2010].

30 Vgl. hierzu Pohl/Ostrowski (2010).

31 Wie bereits erwähnt sollte eine Open-Data-Plattform denen, die sich den Betrieb nicht leisten können, dennoch als Service angeboten werden.

32 Vgl. zu diesen Überlegungen Birbeck, Mark (2009): RDFa and Linked Data in UK government web-sites. Einsehbar unter http://blogs.talis.com/nodalities/2009/07/rdfa-and-linked-data-in-uk-government-web-sites.php [05.08.2010].

33 Vgl. Tennison, Jeni (2009): Your Website is Your API: Quick Wins for Government Data. Einsehbar unter http://www.jenitennison.com/blog/node/100 [05.08.2010].

34 Vgl. Pohl/Ostrowski (2010), Abschnitte 2 und 3, wo wir die Ziele und den Nutzen von Linked Data erläutern.

35 Vgl. Pohl/Ostrowski (2010), Abschnitt 6.3.

36 Auch andere Ansätze zur Beschreibung von Datensammlungen sollten näher untersucht werden. Da Bibliothekskataloge häufig in Form mehrerer Dateien herausgegeben werden (vgl. etwa die freien Katalogdaten der USB Köln unter http://opendata.ub.uni-koeln.de/dumps/, [05.08.2010]) kommt z.B. auch die Verwendung von OAI-ORE in Frage. OAI-ORE steht für Open Archives Initiative - Object Reuse and Exchange, siehe http://www.openarchives.org/ore/ [05.08.2010].

Data und Freien Katalogdaten. Da dcat das Potential hat, ein weit verbreiteter Standard zu werden, sollten möglichst viel aus diesem Vokabular für unsere Zwecke nachgenutzt werden.
Grundlegende dcat-Konzepte sind "Catalog", "CatalogRecord" und "Dataset", die wie folgt miteinander in Beziehung stehen:[37]

- Das oberste Konzept ist ein dcat-Catalog.
- Ein dcat-Catalog setzt sich aus einzelnen dcat-CatalogRecords zusammen
- Die einzelnen dcat-CatalogRecords beschreiben Datasets, in unserem Falle wären dies Abzüge von Bibliothekskatalogen (bzw. Auszüge daraus).

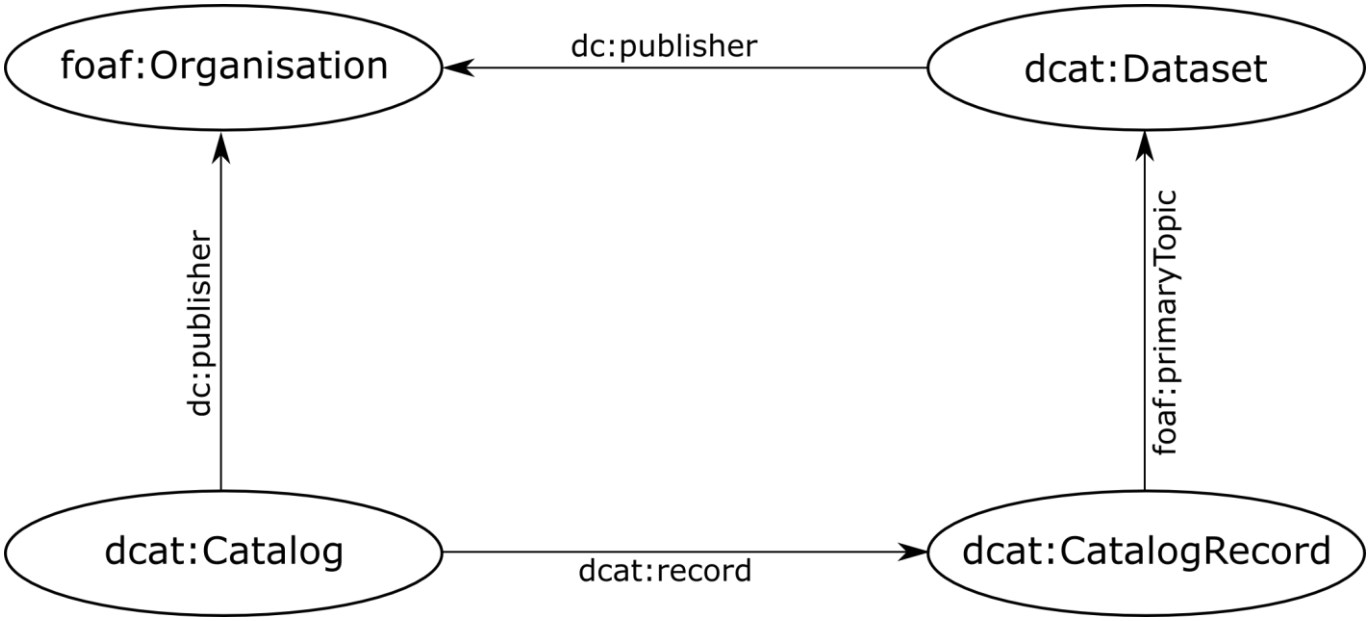

Abbildung 1: Vereinfachte Darstellung des dcat-Modells

Wir haben versuchsweise einige der freigegebenen bibliographischen Datensammlungen mit dcat beschrieben. Dabei traten Probleme des aktuellen Entwurfs wie auch Unzulänglichkeiten der Anwendung auf Open Bibliographic Data zutage. Unsere Überlegungen und Probleme mit dem Vokabular werden als Feedback an die W3C-Gruppe Data Catalog Vocabulary Project gehen. Wir werden in der nächsten Zeit eine entsprechende Dokumentation sowie unser Feedback veröffentlichen.[38]

---

37 Wir sehen hier vom Konzept der "Distribution" ab, weil dieses noch sehr unklar definiert ist und insbesondere dieser Teil des Vokabulars noch einigen Veränderungen unterliegt. Vgl. den Issue-Log der Arbeitsgruppe unter http://www.w3.org/egov/IG/track/products/19 [05.08.2010].

38 Im Fortgang unserer Experimente mit dcat wurde uns immer klarer, dass die Beschreibung der freigegebenen Bibliothekskataloge nur ein Schritt in Richtung des Semantic Web sein kann. Die Umsetzung von Linked-Data-Prinzipien in der Bibliothekswelt und damit der Beitrag der Institutionen zum Semantic Web kann nicht bei der Beschreibung freier Katalogdaten stehenbleiben. Vielmehr ist es sinnvoll und zukunftsträchtig darüber hinaus auch und gerade Institutionen, Sammlungen und Kataloge im allgemeinen in RDF zu beschreiben und diese Beschreibungen zu aggregieren. Auf dieser Ebene lassen sich sehr nützliche Dienste für Recherchierende denken, die nach Findmitteln, Sammlungen und Institutionen für bestimmte Fachbereiche suchen. Im hbz wurden bereits erste Beschreibungen von bibliothekarischen Institutionen in RDF erzeugt (vgl. Ostrowski, Felix (2010): Building a Linked Data based index of library institutions. Einsehbar unter http://blog.lobid.org/2010/07/building-linked-data-based-index-of.html, [05.08.2010]). Zum jetzigen Zeitpunkt und in diesem Rahmen scheint uns eine detailliertere Erläuterung der Aktivitäten auf diesem Gebiet nicht angebracht.

4.3 Versionierung

- Neben dem Beschreiben von Rohdaten ist die Verwaltung der Daten über die Zeit eine Kernanforderung. Wie bereits erwähnt, ist der einfachste Ansatz, in regelmäßigen Abständen neue, aktualisierte Gesamtabzüge der Daten bereitzustellen. Dieser bringt jedoch einige Schwierigkeiten mit sich:
- Es ist alles andere als trivial, aktualisierte Daten der ursprünglichen Quelle mit einer lokalen Version der Daten zusammenzuführen.
- Auch andersherum ist es aufwändig, Änderungen, die an verschiedenen Stellen stattgefunden haben, in die ursprünglichen Daten zurückfließen zu lassen.
- Um auch auf alte Versionen zugreifen zu können, müssen alle Komplettabzüge langfristig vorgehalten werden.

Diese Grundprobleme sind sowohl in der Softwareentwicklung als auch in kollaborativen Schreibumgebungen wie Wikis wohlbekannt. Ihnen wurde mit dem Konzept der Versionskontrolle begegnet. Klassische Versionierungstools sind dazu ausgelegt, Änderungen an Texten zu verwalten. Die Änderungsverfolgung findet auf Zeilenebene statt; sie sind im allgemeinen nicht darauf optimiert, binäre Daten - wie etwa im MAB-Format - zu versionieren.

Was bedeutet dies nun hinsichtlich ihrer Nachnutzbarkeit für die Versionierung von bibliographischen Daten? Es wird deutlich, dass es nötig ist, die Daten statt in einem binären Format in einem Textformat[39] zu exportieren, idealerweise UTF-8 codiert. Ob es günstiger ist, einen Datenexport in einer großen Datei oder in vielen kleinen Dateien zu versionieren, ist noch nicht abschließend geklärt.

```
0331.001:Semantic Web
0335.001:Grundlagen
0359.001:Pascal Hitzler ...z.B. laufende Projekte (dcat, CKAN)
0403.001:1. Aufl.
0410.001:Berlin [u.a.]
0412.001:Springer
0425.001:2008
0433.001:X, 277 S. : graph. Darst.
0451.001:eXamen.press
0540.001:978-3-540-33993-9
0540.002:3-540-33993-0
0542.001:kart. : EUR 24.95 (DE), EUR 25.70 (AT), sfr 41.00
```

Tabelle 1: Bibliographische Daten der USB Köln in einem Zeilenbasierte Format

39 Es sei an dieser Stelle angemerkt, dass es mit N-Triples (http://www.w3.org/TR/rdf-testcases/#ntriples [05.08.2010]) auch für künftig als RDF publizierte bibliographische Daten ein für diese Art der Versionierung geeignetes Format gibt. Es handelt sich dabei um ein Format, in dem jedes Tripel auf einer eigenen Zeile komplett ausgeschrieben wird. Zu den Vorzügen dieses Formats gehört sicherlich nicht die Menschenlesbarkeit, sondern vielmehr die Möglichkeit mit vielen Standardtools darüber operieren zu können, vgl. Bendiken, Arto (2010): RDF for Intrepid Unix Hackers: Grepping N-Triples. Einsehbar unter http://blog.datagraph.org/2010/03/grepping-ntriples [05.08.2010].

```
diff --git a/sample.txt b/sample.txt
index 5db3d78..bb494a0 100644
--- a/sample.txt
+++ b/sample.txt
@@ -10,3 +10,7 @@
 0540.001:978-3-540-33993-9
 0540.002:3-540-33993-0
 0542.001:kart. : EUR 24.95 (DE), EUR 25.70 (AT), sfr 41.00
+0662.001:http://digitool.hbz-nrw.de:1801/webclient/DeliveryManager?pid=2415472
+0662.002:http://digitool.hbz-nrw.de:1801/webclient/DeliveryManager?pid=2220100
+0663.001:Link-Text: Semantic Web; Interna: Verlagsdaten Springer
+0663.002:Link-Text: Semantic Web; Interna: Inhaltsverzeichnis
```

Tabelle 2: Eine Änderung (diff), die obige Datei um weitere Informationen ergänzt

Auch bei der Versionsverwaltung stellt sich die Frage, wie diese organisiert wird. Der grundlegende Unterschied besteht hier wieder zwischen zentralen und dezentralen bzw. verteilten Lösungen[40]. Bei der zentralen Versionsverwaltung gilt das klassische Client-Server-Paradigma. Es gibt einen zentralen Server, der die Versionshistorie verwaltet und von dem man die jeweiligen Daten beziehen kann, um sie lokal zu bearbeiten. Änderungen können nur an diesen Server zurückgespielt werden. Die verteilte Versionsverwaltung hingegen lässt sich als peer-to-peer-Ansatz beschreiben. Das heißt, das jede Kopie des Repositories insofern eigenständig ist, als dass sie die gesamte Versionshistorie enthält. Dies bedeutet, dass sie (1) auf eine Anbindung an einen zentralen Server nicht weiter angewiesen ist und dass (2) jede Kopie als Vorlage für weitere Kopien dienen kann.

Es gibt verschiedene Gründe[41], die für die Wahl eines verteilten Versionsverwaltungssystems sprechen. Zum einen bietet ein solches System die einfache Möglichkeit, spontan kollaborativ zu arbeiten, ohne ein zentrales Repository damit zu belasten. Desweiteren ergibt sich wieder eine größere Unabhängigkeit, denn es gibt keinen zentralen Server, dessen Ausfall oder Nichterreichbarkeit sich negativ auf das Arbeiten mit den Daten auswirken könnte. Da, wie bereits erwähnt, in einem verteilten System jede Kopie eigenständig ist, wird in diesem Szenario darüber hinaus auch nach dem „Lots of copies keep stuff safe“-Prinzip nicht nur der jeweilige Datenbestand, sondern auch die gesamte Versionsgeschichte dezentral archiviert.

5 Fazit und Ausblick

Dieser Text hat die konzeptionellen Grundlinien einer Open-Data-Infrastruktur diskutiert. Es ist klar geworden, dass es sich um eine facettenreiche Problematik handelt. Die sich daraus ergebenden Fragen reichen von der Publikation über die Beschreibung bis hin zur Änderungsverwaltung. Eine Komplettlösung für sämtliche Aspekte gibt es bislang nicht, viele Fragen sind noch unbeantwortet. Allerdings gibt es für die einzelnen Facetten bereits vielversprechende Anknüpfungspunkte, z.B. in

40 Eine weitverbreitete Software für die zentrale Versionsverwaltung ist Subversion (http://subversion.apache.org/), für die verteilte Versionsverwaltung kann z.B. git (http://git-scm.com/) verwendet werden.

41 Vgl. Pollock, Rufus (2010b): We Need Distributed Revision/Version Control for Data. Einsehbar unter http://blog.okfn.org/2010/07/12/we-need-distributed-revisionversion-control-for-data/ [05.08.2010].

Gestalt von Versionsverwaltungstools aus der Open-Source-Community oder in Form von laufenden Projekten wie dcat und CKAN.

Dass hauptsächlich Akteure außerhalb des Bibliothekswesens an der Lösung der betrachteten Probleme arbeiten, demonstriert eindrücklich die Chancen, die sich aus einer Vernetzung und Kooperation mit jenen ergeben. Das Wissen von Bibliothekarinnen und Bibliothekaren über die Beschreibung von Ressourcen und die Erstellung von Katalogen und die Erfahrungen etwa des W3C oder der OKFN in Bezug auf Webstandards und Open Data versprechen eine starke gegenseitige Befruchtung.

Jede Infrastruktur hat wenig Wert, wenn sie nicht benutzt wird. Dies gilt auch für eine Open-Data-Infrastruktur. Neben den hier erläuterten technischen Fragen gilt es auch und vor allem eine Open-Data-Praxis zu etablieren. Die Erfahrungen mit Open-Access-Repositorien haben gezeigt, dass man sich auf eine "Build it and they will come"-Strategie nicht verlassen kann.[42] Vielmehr handelt es sich hierbei um die größte Herausforderung: das Bewusstsein für die Notwendigkeit einer Open-Data-Praxis zu schaffen und diese durch eine entsprechende Infrastruktur zu unterstützen.

6 Quellen

Bendiken, Arto (2010): RDF for Intrepid Unix Hackers: Grepping N-Triples. Einsehbar unter http://blog.datagraph.org/2010/03/grepping-ntriples [05.08.2010].

Birbeck, Mark (2009): RDFa and Linked Data in UK government web-sites. Einsehbar unter http://blogs.talis.com/nodalities/2009/07/rdfa-and-linked-data-in-uk-government-web-sites.php [05.08.2010].

CERN (2010): The CERN Library publishes its book catalog as Open Data. Einsehbar unter http://library.web.cern.ch/library/Library/announcement.html [05.08.2010].

Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (2010): Freigabe der Katalogdaten: Kölner Bibliotheken leisten Pionierarbeit. Pressemitteilung, einsehbar unter: http://www.hbz-nrw.de/dokumentencenter/presse/pm/datenfreigabe [05.08.2010].

Ostrowski, Felix (2010): Building a Linked Data based index of library institutions. Einsehbar unter http://blog.lobid.org/2010/07/building-linked-data-based-index-of.html [05.08.2010].

Pohl, Adrian / Ostrowski, Felix (2010): "Linked Data" - und warum wir uns im hbz-Verbund damit beschäftigen. Erscheint in B.I.T. Online 3/2010. Einsehbar u.a. unter http://www.hbz-nrw.de/dokumentencenter/produkte/lod/ [05.08.2010].

42 Die Annahme, dass die Bereitstellung eines Repositorien-Dienstes gleichsam automatisch zu dessen Akzeptanz und Nutzung führt, hat sich als allzu optimistisch erwiesen. Vgl. Salo, Dorothea (2008): Innkeeper at the Roach Motel, Library Trends 57:2. Pre- und Postprint einsehbar unter http://minds.wisconsin.edu/handle/1793/22088 [05.08.2010].

Pohl, Adrian (2010): Open Data im hbz-Verbund. Erscheint in ProLibris 3/2010. Preprint einsehbar unter http://www.hbz-nrw.de/dokumentencenter/produkte/lod/ [05.08.2010].

Pollock, Rufus (2006): The Four Principles of (Open) Knowledge Development. Einsehbar unter http://blog.okfn.org/2006/05/09/the-four-principles-of-open-knowledge-development/ [05.08.2010].

Pollock, Rufus (2010a): ORDF - the OKFN RDF Library. Einsehbar unter http://blog.okfn.org/2010/07/02/ordf-the-okfn-rdf-library/ [05.08.2010].

Pollock, Rufus (2010b): We Need Distributed Revision/Version Control for Data. Einsehbar unter http://blog.okfn.org/2010/07/12/we-need-distributed-revisionversion-control-for-data/ [05.08.2010].

Salo, Dorothea (2008): Innkeeper at the Roach Motel, Library Trends 57:2. Pre- und Postprint einsehbar unter http://minds.wisconsin.edu/handle/1793/22088 [05.08.2010].

Tennison, Jeni (2009): Your Website is Your API: Quick Wins for Government Data. Einsehbar unter http://www.jenitennison.com/blog/node/100 [05.08.2010].

Torkington, Nat (2010): Truly Open Data. Einsehbar unter http://radar.oreilly.com/2010/03/truly-open-data.html [05.08.2010].