ZBW Publikationsarchiv

Publikationen von Beschäftigten der ZBW – Leibniz-Informationszentrum Wirtschaft
Publications by ZBW – Leibniz Information Centre for Economics staff members

Neubert, Joachim

Article

Linked Open Data und die Bibliothekspraxis

Zeitschrift für Bibliothekswesen und Bibliographie (ZfBB)

Suggested Citation: Neubert, Joachim (2014) : Linked Open Data und die Bibliothekspraxis, Zeitschrift für Bibliothekswesen und Bibliographie (ZfBB), ISSN 1864-2950, Klostermann, Frankfurt a.M., Vol. 61, Iss. 2, pp. 59-67,
https://doi.org/10.3196/186429501461217

This Version is available at:
http://hdl.handle.net/11108/157

Kontakt/Contact
ZBW – Leibniz-Informationszentrum Wirtschaft/Leibniz Information Centre for Economics
Düsternbrooker Weg 120
24105 Kiel (Germany)
E-Mail: info@zbw.eu
https://www.zbw.eu/de/ueber-uns/profil-der-zbw/veroeffentlichungen-zbw

Standard-Nutzungsbedingungen:
Dieses Dokument darf zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden. Sie dürfen dieses Dokument nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen. Sofern für das Dokument eine Open-Content-Lizenz verwendet wurde, so gelten abweichend von diesen Nutzungsbedingungen die in der Lizenz gewährten Nutzungsrechte.

Terms of use:
This document may be saved and copied for your personal and scholarly purposes. You are not to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public. If the document is made available under a Creative Commons Licence you may exercise further usage rights as specified in the licence.

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Mitglied der Leibniz-Gemeinschaft

Linked Open Data und die Bibliothekspraxis

Joachim Neubert

Zeitschrift für Bibliothekswesen und Bibliographie, Jg. 61, Heft 2 (2014) – Preprint

Um 2008 begannen Bibliotheken sich hierzulande für Linked Open Data zu interessieren. Die „Semantic Web in Libraries“ Konferenz SWIB (s. Bericht in diesem Heft) startete 2009 als „Fachtagung für innovative Bibliothekare“ – mit seitdem stets steigender Beteiligung. Wenn das Thema auch inzwischen etabliert ist – ist es für die Praxis in Bibliotheken relevant geworden?

Dieser Artikel demonstriert an Beispielen das Potential von Linked Open Data für *Applikationen mit praktischem Mehrwert für BibliotheksnutzerInnen*. Er geht dann auf *Datensets und Services* ein, die in den letzten Jahren als Bausteine für solche Applikationen verfügbar geworden sind. Und schließlich wird er Entwicklungen skizzieren, die Linked Open Data *stärker ins Zentrum der (IT-) Landschaft* von Bibliotheken rücken.

Linked Open Data – das sind zunächst einmal Daten. Diese Daten können über ihre eindeutigen Webadressen (http://...) referenziert und damit praktisch unbegrenzt miteinander verknüpft werden. Wie das World Wide Web gezeigt hat, liegt in dieser einfachen und unbegrenzten Verknüpfbarkeit ein ungeheures Potential. Und zugleich ist sie das, was diese Art von Daten unübersichtlich macht. Die traditionellen Record-Strukturen und -Grenzen existieren in der „Linked-Data-Cloud“ nicht mehr. Auf Metadatenebene entstehen Verbindungen zwischen Domänen und Einrichtungen, die bislang kaum Berührung miteinander hatten. Generische Linked-Data-Browser, die alle diese Verbindungen ohne Gewichtung auflisten, vermitteln einen Eindruck von dieser Welt der atomisierten Zusammenhänge. Für die spielerische Erkundung unbekannter Datensets nach dem Grundsatz „follow your nose“[1] können solche Browser für technisch orientierte AnwenderInnen und Informationsspezialisten nützlich sein. EndnutzerInnen kommen mit Linked Data jedoch in aller Regel indirekt in Kontakt: durch Anwendungen, die die vernetzten Daten funktional einbinden.

Welche technische Lösung hierbei im Hintergrund zum Einsatz kommt macht für die Funktionalität an der Oberfläche in vielen Anwendungsfällen kaum einen Unterschied. Klassische Datenabzüge und –importe können genauso zum Ziel führen, wie Mashup-APIs der Web-2.0-Generation oder eben die in der Linked-Data-Welt genutzten Semantic-Web-Techniken. Dagegen können die Unterschiede für die Erstellung und die Aufwände für Pflege und Wartung der Anwendungen und ihrer jeweiligen Datenbasen erheblich sein – auch mit Blick auf die Kosten. Und beim letzten Punkt schließt sich der Kreis: Alles, was solche Anwendungen bieten, selbst zu machen würde häufig so teuer, dass es gar nicht in Frage käme – und damit auch für die Nutzerinnen und Nutzer nicht verfügbar wäre.

Linked Data nutzen

Externe Datensets in bibliothekarische Anwendungen einbinden

Die „Pressemappe 20. Jahrhundert“[2] (P20) demonstriert so einen Anwendungsfall. Zeitlich deckt das weltweit einzigartige, von der ZBW (Zentralbibibliothek für Wirtschaftswissenschaften, bzw. den Vorgängerorganisationen) erstellte und öffentlich zugängliche Pressearchiv fast das gesamte 20. Jahrhundert ab; inhaltlich – mit über 20 Millionen Dokumenten – praktisch alles, was in dieser Zeit wirtschaftlich, aber auch politisch aktuell war. Die LOD-Anwendung stellt Dossiers des Personen- und Firmenarchivs zur Verfügung und befindet sich derzeit im Beta-Stadium[3]. Die Dossiers bestehen jeweils aus z.T. Hunderten von eingescannten Presseartikeln, manchmal auch weiteren Materialien wie etwa Geschäftsberichten. Jedes Dossier, jedes Dokument und jede Seite erhält eine feste Adresse im Web und ist damit eindeutig zitierbar. Die Orientierung im Archiv ist jedoch zunächst einmal schwierig. Im „Rohzustand“ waren die Mappen lediglich mit dem Namen sowie Geburts- und Sterbejahr der entsprechenden Person beziehungsweise Namen und Sitz der Firma beschriftet. Im Zuge der Digitalisierung wurde bereits vorausschauend die PND-(heute GND-)Nummer der Personen zusätzlich erfasst, die GKD/GND-Nummern von Firmen werden derzeit nachgetragen.

Mit der Veröffentlichung der Personennormdatei (bzw. seit 2012 der Gemeinsamen Normdatei) als Linked Open Data[4] ließen sich die gespeicherten PND-IDs in Webadressen (URIs) umsetzen, die im ersten Schritt auf Literatur von und über diese Personen bei der DNB verlinken. Zugleich erschließt sich damit jedoch auch die „Linked-Data-Cloud“: Über VIAF, das Virtual International Authority File, das die Normdatenbestände von über 30 Nationalbibliotheken und anderen großen Anbietern zusammenführt[5], sind auch URIs in diese Bestände hinein zugänglich. Und damit nicht genug: Ein gemeinsames Projekt von DNB und Wikipedia legte den Grundstein für die Verknüpfung von bibliothekarischen Normdaten und Wikipedia-Einträgen, die inzwischen von tausenden von Freiwilligen fortgeschrieben und auf Stand gehalten wird. Am Fuß zahlreicher Wikipediaseiten findet sich dazu eine Linkleiste „Normdaten (Person)“, inzwischen gelegentlich auch „Normdaten (Körperschaft)“. Da diese Verknüpfungen auch von GND und VIAF aus zugänglich sind, kann die Pressemappen-Applikation im zweiten Schritt auf Daten des Wikipedia-Linked-Data-Extrakts DBpedia[6] zugreifen.

Das ermöglicht es, die Namen auf den Mappen in einen Kontext zu setzen: Zu Ernst Abbe (1840-1905) etwa erfährt man aus dem Wikipedia/DBpedia-Abstract, dass er als Physiker die Grundlagen der modernen Optik schuf und für die Firma Carl Zeiss unternehmerisch wirkte. Ein Bild macht die Person anschaulich, ein Link führt zum zugehörigen Wikipedia-Artikel mit umfassenden Informationen. Die Aufwände, eigene Abstract-Texte neu zu verfassen und urheberrechtlich freie Bilder zu recherchieren wären für eine solche Anwendung prohibitiv hoch. Das gilt erst recht für die englischsprachige Version der Seiten: Während in Wikipedia/DBpedia Links zur englischen Fassung in aller Regel verfügbar sind (häufig auch zu zahlreichen anderen Sprachen), wäre eine intellektuelle Übersetzung auf mehrere tausend Personendossiers bezogen finanziell „nicht darstellbar“.

Die Aufwände für die Programmierung von Zugriffen im „Web of Data“ sind dagegen relativ gering – und völlig unabhängig von der Anzahl der Datensätze, für die sie genutzt werden. Zudem lassen sich Erfahrungen von der einen auf die andere Datenquelle übertragen – vor allem wenn die Quellen Standard-(RDF-) Vokabulare nutzen und eingeführten Linked-Data-Patterns folgen.

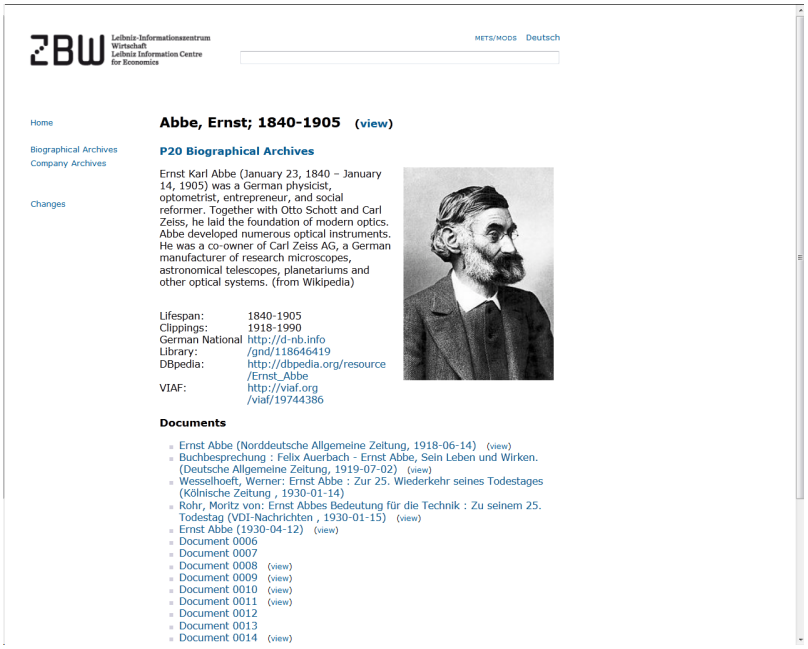

Abb. 1 ZBW-Pressearchiv 20. Jahrhundert (beta): Bilder und multilinguale Beschreibungen per LOD

Neben der verbalen Information in Texten kann eine Anwendung auch Strukturinformation aus dem Datenweb übernehmen. Wieder ein Beispiel aus der „Pressemappe 20. Jahrhundert“: Die Mappentitel des Firmenarchivs enthalten neben dem Firmennamen in Klammern den Firmensitz. Diese Angaben können extrahiert und mit Städtenamen des GeoNames-Service[7] gematcht werden. Etwa für „Frankfurt am Main“ wird ein eindeutiger URI[8] zurückgeliefert. Unter diesem URI sind bei GeoNames u.a. die Geolokation (Länge/Breite) und die administrative Zugehörigkeit (Hessen -> Deutschland) zu erfahren. Die Anwendung nutzt diese Daten nicht nur, um die Firmen auf einer Karte zu lokalisieren. Sie erzeugt daraus auch die Menüstruktur „Firmen nach Ländern“ -> „Firmen mit Sitz in Frankfurt am Main“ – das heißt, sie „implantiert“ die geografische Navigation aus der externen Linked-Data-Quelle GeoNames in die Applikation.

Die Möglichkeiten sind damit längst nicht ausgereizt: Da GeoNames auch Abfragen „im Umkreis von x km“ zulässt, könnten bei einer unscharfen Geo-Suche auch Firmen in Eschborn oder Offenbach mit eingeschlossen werden. Eine zusätzliche Menüebene „Firmen nach Bundesländern“ wäre ebenfalls leicht implementierbar. Selbst ausführliche Informationen über die einzelnen Städte oder Ortschaften, an denen die Firmen ihren Sitz hatten, sind abrufbar und integrierbar – denn GeoNames veröffentlicht auch DBpedia-Links.

Was an zusätzlichem Kontext für die Benutzung bereitgestellt werden soll, wird so zur Designentscheidung für das „User Interface“: Macht es im Rahmen dieser Applikation Sinn, oder würde eher „information overflow“ generiert? Die mageren Ausgangsmetadaten intellektuell entsprechend anzureichern und zu verlinken wäre jedenfalls viel zu teuer – die implementierte Komfortfunktion Geo-Navigation würde ohne Linked Open Data schlicht nicht existieren.

Mit der Verfügbarkeit von offenen Daten - sowohl im RDF-Format als auch in weniger ausdrucksstarken Formaten wie simplen Text-Tabellen (CSV) – entstand ein Anreiz, mit solchen Daten zu experimentieren

und Software dafür zu entwickeln. Wenn solche Tools nicht eng auf ein einzelnes Projekt beschränkt sind, stehen sie häufig als Open Source Software zur Verfügung. So etwa das „europeana4D interface“[9], das einen geographisch-räumlichen und zugleich zeitlich strukturierten Blick auf Daten erlaubt. Durch Einsatz aktueller Browser-Technologie können BenutzerInnen interaktiv z.B. in Europakarten zoomen, auf denen zugleich mit der Änderung des zeitlichen Bezugs sich die Landesgrenzen ändern und die Ereignisse der jeweiligen Zeit aufscheinen[10]. Falls es also für eine Anwendung wie die Pressemappe angemessen erschiene, Daten zur zeitlichen Existenzspanne ihrer Objekte abzufragen (für Firmen z.B. aus der GKD/GND und/oder DBpedia), könnten diese mit Hilfe eines fertigen Programmbausteins korrekt auf historische Landkarten projiziert werden.

Mit den Abhängigkeiten zwischen verteilten Datenquellen umgehen

Der am obigen Beispiel beschriebene Online-Zugriff auf das „Web of Data“ zur Laufzeit der Anwendung hat viele Vorteile: Eine eigene Datenhaltung fällt weg, ebenso wie die häufig mühsamen Prozesse zum Import/Export der Daten. Wenn, wie etwa bei der an der UB Mannheim entwickelten Mashup-Anwendung zur Unterstützung der bibliotheksinternen Literaturauswahl und Erwerbung[11], zahlreiche unterschiedliche Quelldatenbestände herangezogen werden, würde deren Replizierung allein schon prohibitiv hohe Aufwände generieren. Vor allem aber erfolgt bei Online-Zugriff die Aktualisierung der Daten (durch neue Erkenntnisse, oder für zusätzliche Personen oder Firmen, deren Dossiers online verfügbar werden) automatisch. Allerdings kommt damit zugleich ein Element der Abhängigkeit von externen Anwendungen ins Spiel: Wenn etwa GND und/oder DBpedia zeitweilig nicht verfügbar sind, fehlen die entsprechenden Informationen. Im Sinne lose gekoppelter Systeme – ein Programmierparadigma, das gerade im Linked-Data-Kontext elementar ist – sollte in so einem Fall das System ohne die anreichernden Informationen weiterlaufen. Die zunehmende Vernetzung der Daten im Web lässt manchmal auch Alternativen zu: Als zum Beispiel nach der GND-Umstellung im April 2012 die DBpedia-Links von der DNB für längere Zeit nicht mehr ausgeliefert wurden, konnte die Pressemappen-Anwendung dies umgehen, indem sie stattdessen einen VIAF-Service nutzte, der zu einer ID alle bekannten Links zu anderen Daten liefert („justlinks“). Kurzfristige Aussetzer können gegebenenfalls durch einen Standardmechanismus im Web, das „Caching“ in temporären Zwischenspeichern, überbrückt werden. Bei stabileren Daten und/oder einer ausgeprägteren Abhängigkeit, wie am obigen Beispiel der Geolokationen, kann ein solches Caching auch in „eigenen“ Datenspeichern sinnvoll sein – mit einer vollständigen initialen Datenübernahme und einer Aktualisierung durch „Feeds“ mit laufenden Änderungen.

Der oben angeführte „justlinks“-Service benutzt übrigens keine Semantic-Web-Technologie, sondern liefert alle bekannten Links zu einer Person im einfach zu verarbeitenden JSON-Format. Nicht der „reinen Lehre“ des Semantic Web folgt auch eine andere Anwendung, die im Rahmen der oben erwähnten Kooperation zwischen DNB und Wikipedia entstand – und doch ein herausragendes Beispiel für die Mächtigkeit der Grundidee von Linked Data ist: Die „Wikipedia-Personensuche“, ebenfalls in der Normdaten-Linkleiste der deutschsprachigen Wikipedia verlinkt. Für „Karl Marx“ etwa gelangt man auf eine Seite des Wikipedia-Toolservers mit Links zu zahlreichen Bibliotheksbeständen, aber auch zu Volltexten, Nachlassarchiven, Bilddatenbanken und sogar Musikstücken[12]. Ermöglichst wird das durch einen einfachen Mechanismus: Die einzelnen teilnehmenden Institutionen publizieren eine Liste von GND-Personen-IDs, und zugleich die „Bauanleitung“ für einen Link, mit dem in ihrem System auf

Informationen zur entsprechenden Person-ID zuzugreifen ist. Der Toolserver generiert aus diesen Informationen Links auf der Personenseite. Der BEACON[13] genannte Standard ist so einfach umzusetzen, dass sich inzwischen mehrere hundert Institutionen mit dezentral erstellten Linkdateien an dieser „Grassroot-Vernetzung“ beteiligen.

So oder so: Benutzerinnen und Benutzer profitieren von der Anreicherung der Daten. Der spezifische Semantic-Web-/Linked-Data-Hintergrund der Anwendung ist für sie in aller Regel nicht relevant, und oft auch gar nicht erkennbar - „… as it should be“, wie OCLC’s „Technology Evangelist“ Richard Wallis im Hinblick auf ein neues Linked-Data-Angebot der britischen BBC anmerkt. „If you can see the technology, you have got it wrong.”[14]

Fallbeispiel BBC: Linked-Data-basierte Anwendungen im Produktionseinsatz

Die BBC setzt Linked Open Data inzwischen bei mehreren Angeboten produktiv ein und experimentiert bei weiteren damit.[15] Für die täglich 1000-1500 Sendungen wäre es unmöglich, begleitende Webseiten intellektuell zu erstellen. Für BBC Music, den „Klassiker“ unter den BBC-LOD-Aktivitäten, stellt deshalb der „MusicBrainz“-LOD-Dienst das „navigational backbone“ bereit. Er liefert Identifier (URIs) für Musiker, Bands und Musikstücke – und deren Verknüpfungen untereinander. Aus Wikipedia werden Kurzbiographien übernommen. Auf der BBC-Webseite zu Norah Jones[16] sind nicht nur Stücke der Sängerin und Sendungen über sie abrufbar, sondern auch BBC Rezensionen ihrer Stücke, eine Kurzbiographie aus Wikipedia, Links auf BBC-Seiten zu den Gruppen, in denen sie gespielt hat, und auch zu ihrem Vater Ravi Shankar. Externe Links führen auf Youtube, Discogs, auf ihre Homepage und in die internationale Filmdatenbank. Die Herkunft der inhaltlichen und Strukturdaten ist auf der Seite jeweils mit angegeben – womit die BBC einerseits die Urheber würdigt, und zugleich andererseits deutlich macht, welche Ergänzungen von Dritten stammen – und wofür sie selbst inhaltlich die volle Verantwortung übernimmt. Die Daten aus dem Web werden scheinbar wie von selbst, nämlich durch User der eingebundenen Dienste, aktualisiert, wenn etwa ein Künstler stirbt, oder ein neues Album einer Band veröffentlicht wird. Solche Pflegeaufwände könnte – für die Breite des Angebots – auch eine große Institution wie die BBC nicht aufbringen.

Der BBC Wildlife Finder[17] ist wohl das bunteste Linked-Data-getriebene Angebot weit und breit. DBpedia wurde hier als kontrolliertes Vokabular zur Verschlagwortung von Programmepisoden, Clips und Nachrichten über Wildtiere eingesetzt.[18] Die detaillierten Beschreibungen und Fotos der einzelnen Gattungen und Arten in Wikipedia unterstützen den Indexierungsprozess. Auch aus anderen Quellen, etwa vom WWF Wildfinder und dem Animal Diversity Web, werden Informationen zusammengetragen – über Linnésche Systematik, Verbreitungsgebiete und Habitate, über das Verhalten und den Bestand, bis hin zu Links auf Progammclips oder Vogelstimmen. Die Daten sollen von anderen Linked-Data-Anwendungen nachgenutzt werden. An deren Entwickler richtet sich die Aussage: „Our website is our API – you can access the site as RDF/XML or RSS.“

An eine andere – und mit über 11 Millionen Besuchern pro Woche sehr viel breitere - Anwenderschicht wendet sich BBC Sport. Seit dem Relaunch der Seite 2012 werden dort Semantic-Web-Technologien im großen Maßstab eingesetzt. Journalisten sollen dadurch mehr Zeit für die Erstellung von Inhalten einsetzen können, und weniger für das Management dieses Contents. „In the past when a journalist wrote a story they would have to place that story on every relevant section of the website. A story about

Arsenal playing Manchester United, for example, would have to be placed manually on the home page, the Football page, the premier league page, the Arsenal page and the Manchester United page – a very time consuming and labour intensive process." Platzierung und Verlinkung können nun automatisiert werden – mit zugleich besseren Navigationsmöglichkeiten für die BenutzerInnen.[19]

Der BBC World Service Archive Prototype kombiniert die Semantic-Web-Techniken mit Textanalyse und einem Crowdsourcing-Ansatz. Das Archiv enthält etwa 70.000 seit 1947 gesendete Programme. Auch hier sind die Metadaten spärlich – daher wurden die digitalisierten Programme automatisch transkribiert. Mit Hilfe von selbst weiterentwickelten und bereits existierenden Linked Data Tools wurden aus Metadaten und transkribierten Texten automatisiert DBpedia-URIs ermittelt, die die Verknüpfungen innerhalb der Wikipedia für das Archiv erschließen[20]. Ihr weiteres Vorgehen schildern die Autoren so: „The resulting interlinks are used to bootstrap search and navigation within this archive and expose it to users. Automated data will never be entirely accurate so we built crowdsourcing mechanisms for users to correct and add data. The resulting crowdsourced data is then used to improve search and navigation within the archive, as well as evaluate and improve our algorithms. As a result of this feedback cycle, the interlinks between our archive and the Semantic Web are continuously improving."[21] In der Prototyp-Anwendung können angemeldete BenutzerInnen per Interface für jeden einzelnen Link ein „upvote"- oder „downvote"-Urteil abgeben oder zusätzliche Wikipedia-Links per „Add Tag"-Dropdownliste auswählen. Nach einem Jahr Laufzeit, im August 2013, hatten sie so 70.000 einzelne Interlinks validiert, invalidiert oder hinzugefügt. Das komfortable Interface und die Tatsache, dass Änderungen sofort sichtbar und für die Suche wirksam werden, dürften wesentlich zur Motivation der BenutzerInnen beigetragen haben. Die positiven Erfahrungen wiederum ermutigten die Macher, auch die Programmtitel und Programmbeschreibungen im Wiki-Stil, mit einsehbarer Versionsgeschichte, editierbar zu machen.

Als jüngster Spross ging BBC News Labs[22] an den Start. Es zielt auf den Kernbereich der Nachrichten – Lokationen, Ereignisse, Politik. Vor allem anhand dieses Projekts wird die „BBC Linked Data Plattform"[23] entwickelt. Die Vielfalt der Anwendungsfelder und –fälle und das fast schon industrielle Ausmaß der Entwicklung und des Einsatzes demonstrieren, wie eine große Institution Linked Data zu einem zentralen Bestandteil ihrer Infrastruktur macht. Für die Triplestores wird eine Verfügbarkeit von 99,99% angegeben. Die Skalierbarkeit wurde von BBC Sport an den Olympischen Spielen 2012 in London erprobt. Und was die Benutzerfreundlichkeit angeht, entsprechen die Anwendungen dem Stand der Anwendererwartungen, die von großen kommerziellen Websites geprägt sind.

Linked Library Data publizieren

Für Bibliotheken mit ihren begrenzten finanziellen und technischen Ressourcen sind solche umfassenden Linked-Data-getriebenen Anwendungslandschaften noch nicht erkennbar. Oft stehen sie noch am Anfang, indem sie zunächst einmal ihre eigenen Daten im Web zugänglich machen. Normdaten, Thesauri und Klassifikationen boten sich dabei als Stammdaten und Referenzpunkte zur Identifizierung von Personen, Institutionen oder Themen – bei gleichzeitig überschaubarem Datenumfang – als Einstiegspunkte an. Die Linked-Data-Publikation der sorgfältig gepflegten kontrollierten Vokabulare von Bibliotheken ist zugleich ein Angebot an die Semantic-Web-Community, diese Daten als Rückgrat für weitergehende Verknüpfungen zu nutzen.[24]

Besonders bei Normdaten ist inzwischen die nationale und internationale Abdeckung gut: Die GND ist nicht nur eigenständig als Linked Open Data im Web veröffentlicht[25], sondern sie ist auch Teil von VIAF. VIAF bildet wiederum eine Grundlage von ISNI, dem ISO Standard Name Identifier (ISO 27729:2012) für öffentliche Identitäten von Personen und Organisationen. ORCID ist ein weiterer komplementärer und mit ISNI koordinierter Ansatz für Identifier aus dem Forschungsbereich, die von den ForscherInnen selbst angemeldet und gepflegt werden können[26].

Für die inhaltliche Erschließung stehen einerseits thematisch umfassende Instrumente wie die deutsche Schlagwortnormdatei (als Teil der GND), die Subject Headings der Library of Congress oder in Teilen die Dewey Decimal Classification bzw. die Universal Decimal Classification als Linked Data zur Verfügung. Fachliche Thesauri wie Eurovoc, Thesoz, Agrovoc, Gemet/Umthes und viele andere sind ebenfalls verfügbar[27]. Als gemeinsames Datenformat hat sich hier weitgehend SKOS (Simple Knowledge Organization System) durchgesetzt. Neben Downloads der Gesamtvokabulare bieten einige Herausgeber SPARQL-Endpoints zur interaktiven Abfrage an. Die Einbettung der RDF-Daten (RDFa) in die produktive Webpräsenz der Thesauri, wie 2009 mit dem Standard-Thesaurus Wirtschaft (STW) erstmals praktiziert, stellt eine weitere Form der Publikation im Semantic Web dar.

Die LOD-Publikation von Titeldaten wurde in Deutschland vor allem von den Verbünden vorangebracht. Den Anfang machte 2010 das hbz, es folgten BVB und KOBV, HeBIS , BSZ sowie die DNB mit der deutschen Nationalbibliographie; beim GBV gibt es eine entsprechende Absichtserklärung. International hat eine ganze Reihe von Nationalbibliotheken ihre Daten als LOD zur Verfügung gestellt. Zumeist wurden große Download-Dateien bereitgestellt, die gelegentlich aktualisiert werden. Von besonderer Bedeutung ist die Entscheidung von OCLC, die weltgrößte bibliografische Datenbank WorldCat mit RDFa anzureichern, so dass die wichtigsten semantischen Informationen aus der Webanzeige ausgelesen werden können. Relevant ist das vor allem für Suchmaschinen, auf die auch das von OCLC verwendete schema.org Vokabular zielt. Für bibliothekarischen Datenaustausch, oder auch nur für die Wiederverwendung in Anwendungen anderer Bibliotheken, ist schema.org allerdings zu grobmaschig. Hierfür hat die DINI KIM Titeldaten-AG im letzten Jahr präzisere Empfehlungen[28] veröffentlicht, die bei den Verbünden und der DNB bereits umgesetzt werden. Die W3C Schema Bib Extend Community Group[29] arbeitet parallel an Vorschlägen zu einer Erweiterung des schema.org Vokabulars, insbesondere auch im Hinblick auf Artikeldaten.

Die Publikation der Daten erfolgt inzwischen in der Regel unter freien Lizenzen – sei es die Creative Commons Lizenz CC0, die praktisch auf alle Rechte verzichtet, oder sei es die von WorldCat genutzte Open Data Commons Attribution Lizenz, die eine Herkunftsangabe fordert. Die Verwendung solcher freien Lizenzen ist ein entscheidender Schritt, der die Nachnutzung der Daten in der Praxis erst möglich macht. Erst dadurch konnten vielfältige Linked-Data-Anwendungen entstehen, die ohne aufwendige rechtliche Klärungen die Daten in neuer Weise kombinieren und Mehrwert für die NutzerInnen schaffen. Europeana, die größte europäische Sammlung von Metadaten zum kulturellen Erbe, macht daher CC0 in ihrem „Data Exchange Agreement“ zur verbindlichen rechtlichen Grundlage.

Linked Library Services: Dienste in einer verteilten Welt

Allerdings: Auch freie Daten erfordern, wenn sie lediglich als Gigabyte-große Dumpfiles veröffentlicht werden, eine Menge vorbereitender Arbeit, ehe sie effektiv genutzt werden können. Die

Hardwareanforderungen wie auch das notwendige know-how, um sie in LOD-Datenbanken zu laden und zugänglich zu machen, sind beträchtlich. Auch von den Datenproduzenten bereitgestellte SPARQL-Endpoints, die interaktive Abfragen möglich machen, lösen das Problem nicht: Einerseits ist die Abfragesprache –ähnlich wie SQL – komplex. Eine gute Auswahl von Beispielanfragen, wie sie etwa der Linked-Data-Endpoint der Zeitschrift „Nature“[30] mitliefert, kann den Einstieg in Datenstrukturen und Querysyntax lediglich erleichtern. Andererseits sind SPARQL-Endpoints im Web prinzipbedingt genauso wenig für den Produktiveinsatz geeignet, wie es offene SQL-Schnittstellen wären[31].

Die Culturegraph-Plattform[32], von der DNB betrieben, zielt auf den Aufbau einer allgemein zugänglichen Infrastruktur für solche Daten. Sie versucht die Titeldaten von DNB und Verbünden sowie Normdaten regelmäßig in eine umfassende Datenbank zu laden und miteinander und mit weiteren Daten (insbesondere mit der Wikipedia) zu verknüpfen. Demo-Dienste z.B. zur Kataloganreicherung oder zu erweiterten Personensucheinstiegen - suche „Geheimrat, Weimarer Klassik“, finde „Goethe“ – setzen darauf auf. Forschungsprojekte versuchen, anhand des Gesamtbestandes automatisch Mappings zwischen Klassifikationen abzuleiten oder zahlreiche unterschiedliche Ausgaben zu Werken zu bündeln[33] - letzteres ein verbreitetes Desiderat, das OCLC in WorldCat ebenfalls umzusetzen sucht. Als produktiver Dienst für Endnutzer soll „EntityFacts“ künftig zu allen in Culturegraph bekannten Personen, Körperschaften, Orten, Ereignissen und Titeln ein „Faktenblatt“ mit aufbereiteten Informationen sowie allen bekannten Verknüpfungen zu anderen Entitäten innerhalb von Culturegraph sowie zu externen Informationssystemen liefern.[34]

Einen anderen Ansatz, um die Barrieren bei der Nutzung von Linked Data zu reduzieren, gehen spezialisierte Web Services: Über Programmierschnittstellen (APIs) werden Programmierer von Webanwendungen damit in die Lage versetzt, ohne Hintergrundwissen zu Ontologien, RDF oder SPARQL gezielt Informationen abzufragen. Das hbz bietet mit lobid[35] ein solches API für bibliografische Ressourcen, Organisationen (beide auf Grundlage des Verbundkatalogs), Personen oder Schlagwörter (aus der GND) an. Es ist insbesondere darauf ausgerichtet, performant die Daten für Autosuggest-Auswahllisten zu liefern. Standardmäßig werden die Daten im JSON-Format zurückgegeben, das allen Webprogrammierern bestens bekannt ist (wobei der Linked-Data-spezifische „overhead“ in einen separaten JSON-LD-Kontext verbannt wird).

Einen ähnlichen Weg, jedoch ausgerichtet auf das Fachgebiet Wirtschaftswissenschaften, verfolgt die ZBW mit econ-ws[36]. Autosuggest-Dienste für STW (und Mappings zu anderen Thesauri), JEL-Klassifikation, WirtschaftwissenschaftlerInnen und wirtschaftswissenschaftliche Institutionen stehen hier ebenso zur Verfügung wie Services zur Suchunterstützung, die deutsche und englische Synonyme oder ganze Begriffswolken für das Retrieval bereitstellen, und die in Angeboten wie dem digitalen Repository EconStor auch bereits produktiv genutzt werden.[37]

BIBFRAME, LIBRIS & Co

Die in ihren Auswirkungen auf die Bibliothekswelt sicherlich am weitesten reichende Linked Data Initiative hat 2011 die Library of Congress gestartet: Die „Bibliographic Framework Initiative“ (BIBFRAME) soll eine neue Umgebung für Bibliotheken schaffen, in der „das Netz“ zentral ist und Vernetzung alltäglich wird. Ein neues bibliographisches Datenmodell soll das bisher genutzte MARC21-Format

ablösen (vgl. Abb. 2). Recordstrukturen sollen aufgelöst und durch einen Verbund eigenständiger, sich gegenseitig referenzierender Ressourcen ersetzt werden.

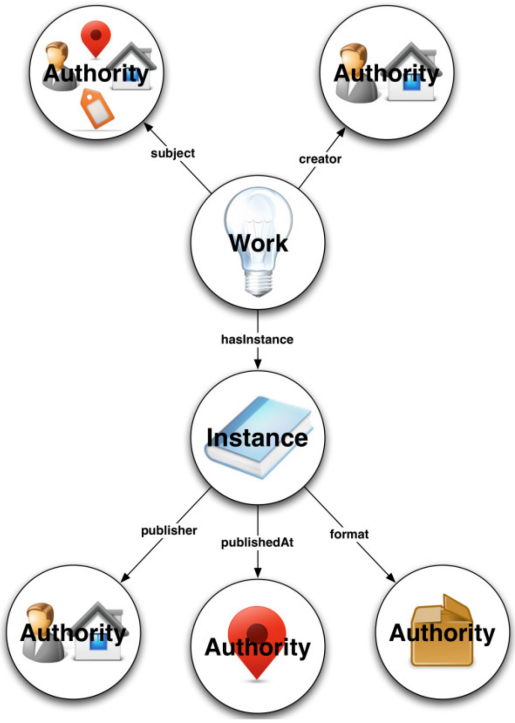

Abb. 2: Grafische Repräsentation des Bibframe-Datenmodells (Quelle: http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf#page=9)

Eine erste Version des BIBFRAME-Vokabulars[38] steht vor der Verabschiedung. BibliothekarInnen, SoftwareentwicklerInnen und Firmen sollen mit diesem Vokabular nun ein Jahr lang experimentieren und es probeweise umsetzen. Auf diesen Erfahrungen aufbauend soll das Vokabular dann noch einmal grundlegend überarbeitet werden. Ernsthafte Implementierungen und Investitionen sollen auf der so erarbeiteten und getesteten „Version 2.0“ des Vokabulars aufsetzen.[39]

Die DNB unterstützt die Initiative und hat auf der SWIB13 eine experimentelle Implementierung des BIBFRAME Datenmodells vorgestellt, die bereits von den Titeldatensätzen im DNB-Portal verlinkt ist.[40] Erste Prototypen eines BIBFRAME Editors, mit dem die LOD-gestützte Eingabe von BIBFRAME-„Datensätzen“ ausprobiert werden kann, werden gerade entwickelt.

Die Anbieter von Bibliothekssystemen halten sich in Sachen Linked Open Data bisher sehr zurück. Die Ex Libris Anwendergruppe IGeLU hat 2012 ein “Manifesto for Linked Open Data in Ex Libris Products“ publiziert. Leider findet die Diskussion darüber in nicht-öffentlichen Kanälen statt; von einem Statement des Ex Libris Managements ist nichts bekannt. Verlautbarte Pläne für PICA scheint es ebenfalls nicht zu geben – obwohl der Hersteller OCLC in Linked-Data-Aktivitäten sehr engagiert ist. Das US-basierte Bibliothekssystem VLTS Open Skies will dagegen (nach Aussage von dessen Präsidenten bei der Library of Congress[41]) das erste System werden, das BIBFRAME unterstützt und das es erlauben soll, im Semantic Web zu navigieren.

Mit OpenCat hat die französische Nationalbibliothek (BnF) zusammen mit der Bibliothek von Fresnes (bei Paris) und einem externen Softwarepartner einen praxisorientierten Prototypen entwickelt[42], der zeigt wie Linked Data für die NutzerInnen öffentlicher Bibliotheken eingebunden werden kann. Bereits zuvor hatte die BnF aus einem Teil ihres Bestandes etwa 100.000 Werkdatensätze (nach FRBR, den Functional Requirements for Bibliographic Records) extrahiert und über data.bnf.fr zusammen mit Personen- und Sach-Normdatensätzen sowie Manifestationsdaten, Digitalisaten etc. unter einer offenen Lizenz als Linked Data bereitgestellt. Diese Daten wurden nun mit den lokalen Bibliotheksdaten kombiniert, so dass bei Suchen im Katalog von Fresnes Personendaten, Illustrationen, digitalisierte Ausgaben aus der BnF und Links in die Wikipedia mit angeboten werden können. Vor allem jedoch konnten die lokalen Exemplardaten anhand der Zuordnung zu Werken strukturiert werden, so dass übersichtliche Trefferlisten die Nutzung erleichtern. Die Recherche wird durch den BnF-Thesaurus unterstützt. Das Beispiel demonstriert, wie auch kleinere Bibliotheken ohne Änderung ihres Bibliothekssystems Linked Open Data in einem auf ihre Daten und Bedürfnisse zugeschnittenen OPAC nutzten könnten.[43]

Am weitesten fortgeschritten bei der Verwendung von Linked Data in der klassischen Bibliothekswelt ist einmal mehr die Schwedische Nationalbibliothek (KB), die zugleich den Verbundkatalog LIBRIS für Universitäts- und öffentliche Bibliotheken betreibt. Sie hatte schon 2008 diesen Katalog als Linked Data publiziert[44], stand aber 2011 vor der Situation, dass sie immer mehr Daten in unterschiedlichen Formaten erhielt und ausgeben wollte – womit die alte Voyager-Software überfordert war : „Obwohl wir über die Jahre das System mit Schicht um Schicht von eigenen Funktionalitäten ummanteln konnten, ist sein Kern, Voyager, nun hoffnungslos veraltet. "[45] Daher traf die KB die mutige Entscheidung, das gesamte Bibliothekssystem durch ein eigenes Entwicklungsteam selbst zu bauen. Das neue, als Open-Source verfügbare System LIBRIS-XL[46] basiert im Kern auf Daten im JSON-LD-Format. Formate wie MARC21 oder MODS werden für die Ausgabe erzeugt, limitieren aber nicht die Funktionalität des Systems selber.

Abb.3: Katalogisierung in LIBRIS-XL (beta)

In der internen Metadatenstruktur sollen - ähnlich wie in BIBFRAME - die Ebenen Werk (entspricht in FRBR grob Work und Expression) und Instanz (entspricht grob Manifestation) benutzt werden. Da Linked Data alle Arten von Daten miteinander in Beziehung setzen könne, verschwänden tendenziell jedoch die Unterschiede zwischen internen und externen Datensets. Protokolle (z.B. Feeds), die normalerweise für externe Datensets benutzt werden, könnten ebenso für interne verwendet werden. Die Kontrolle über die Daten werde von einer technologischen Angelegenheit mehr zur Sache des Vertrauens zwischen Datenproduzenten.[47]

Eine Beta-Version des LIBRIS-Katalogisierungsclients ist bereits online[48]. Die Produktionseinführung und vollständige Ablösung des bisher benutzten Systems soll noch 2014 stattfinden[49].

Fazit

In den letzten Jahren sind große Anstrengungen vor allem in die LOD-Publikation von Bibliotheksdaten geflossen. Der Aufbau von Infrastrukturen, die diese Daten auch effektiv nutzbar machen, und die Entwicklung von darauf aufsetzenden Endnutzeranwendungen stehen – wie ein Vergleich mit der ausdifferenzierten Anwendungslandschaft bei der BBC zeigt – noch am Anfang.

Ein entscheidender Durchbruch ist jedoch erzielt: Die traditionelle Abschottung und Selbstgenügsamkeit der Bibliothekswelt ist durchbrochen worden. Sowohl freie Lizenzen für ihre Daten als auch die Ausweitung des Blicks über die eigenen Datensilos hinaus, auf Standards im Web, die den Austausch mit anderen Communities möglich machen, sind entscheidende Fortschritte, die die Linked-Open-Data-Bewegung gebracht hat. Und allmählich und zunächst punktuell findet dies auch Niederschlag im Kernbereich bibliothekarischer Standards und IT-Systeme.

Wo LOD in Bibliotheken in den produktiven Bereich vordringt, ist es Teil einer Systemumgebung, die konventionellere Technologien selbstverständlich mit einbezieht. Für die BenutzerInnen ist es irrelevant, mit welcher Technologie im Hintergrund gearbeitet wird. Linked Data ist eher Designprinzip als Technologieentscheidung: Ob die Autosuggest-Vorschläge von einem Webservice geliefert wird, der direkt SPARQL-Abfragen kapselt (econ-ws) oder auf Suchmaschinenindices beruht (lobid), macht für die Nutzung keinen Unterschied. Auch ob das Backend auf einer klassischen SQL-Datenbank aufsetzt (data.bnf.fr) oder „reinen“ RDF-Daten (LIBRIS), ist vor allem für Technikexperten spannend.

Für die Bibliotheken sind Semantic-Web-Technologien nicht deshalb revolutionär, weil sie vorher Unmögliches endlich technisch möglich machen würden. Vielmehr hat die Linked-Open-Data-Vision und – noch viel wichtiger – die einfache praktische Verfügbarkeit von entsprechenden Daten eine immense Kreativität freigesetzt. Ideen konnten vergleichsweise leicht in funktionierende Prototypen umgesetzt werden. An den verschiedensten Stellen wurde so demonstriert, was geht, wenn nicht das lokale Bibliothekssystem und auch nicht das jeweils eigene Portal oder der teure Discovery Service, sondern das Web selbst und die Standards, die es erfolgreich gemacht haben, als Plattform angesehen und genutzt werden. In kaum einen anderen Bereich kann Linked Open Data so viel Potenzial freisetzen wie bei den Bibliotheken: Ihre Daten sind schon vom gesellschaftlichen Auftrag her „Open Data“. Der Spardruck, dem sie unterliegen, macht Kooperation und Arbeitsteilung untereinander und mit anderen Communities immer attraktiver. Und so finden, abseits vom Hype,Linked-Data-basierte Anwendungen allmählich Eingang in die tägliche Bibliothekspraxis.

---

[1] Ed Summers, “Following Your Nose to the Web of Data,” *Inkdroid Blog Archive*, January 4, 2008, http://inkdroid.org/journal/2008/01/04/following-your-nose-to-the-web-of-data/.
[2] http://zbw.eu/beta/p20
[3] Joachim Neubert, “The 20th Century Press Archives as Linked Data Application” (presented at the ISWC 2010, Shanghai, 2010).
[4] http://www.dnb.de/lds
[5] http://viaf.org
[6] http://dbpedia.org
[7] http://geonames.org

[8] http://sws.geonames.org/2925533
[9] http://wp1187670.server-he.de/e4d/
[10] Videos unter http://www.youtube.com/watch?v=hVWuSi_6dwY und http://www.youtube.com/watch?v=gFp3tlvDEBc
[11] Philipp Zumstein, "Mash-up for Book Purchasing" (presented at the SWIB13, Hamburg, Germany, November 26, 2013), http://swib.org/swib13/slides/zumstein_swib13_111.pdf.
[12] http://toolserver.org/~apper/pd/person/Karl_Marx
[13] Jakob Voss, Mathias Schindler, and Christian Thiele, "Link Server Aggregation with BEACON," Conference Poster, 2011, http://eprints.rclis.org/handle/10760/15407
[14] Richard Wallis, "BBC Sport Site Built on a Solid Linked Data Foundation," *Data Liberate*, February 3, 2012, http://dataliberate.com/2012/02/bbc-sport-site-built-on-a-solid-linked-data-foundation/.
[15] Yves Raimond et al., "Use of Semantic Web Technologies on the BBC Web Sites," in *Linking Enterprise Data*, Springer 2010
[16] http://www.bbc.co.uk/music/artists/985c709c-7771-4de3-9024-7bda29ebe3f9
[17] http://www.bbc.co.uk/wildlifefinder/
[18] Raimond et al., "Use of Semantic Web Technologies on the BBC Web Sites."
[19] Cait O'Riordan, "Launching the New BBC Sport Website," *BBC Internet Blog*, February 1, 2012, http://www.bbc.co.uk/blogs/bbcinternet/2012/02/launching_bbc_sport_new.html.
[20] Yves Raimond, "Automatically Tagging the World Service Archive," *BBC - Research and Development Blog*, March 20, 2012, http://www.bbc.co.uk/blogs/researchanddevelopment/2012/03/automatically-tagging-the-worl.shtml.
[21] Yves Raimond and Tristan Ferne, "The BBC World Service Archive Prototype," 2013, http://challenge.semanticweb.org/2013/submissions/swc2013_submission_5.pdf.
[22] http://www.bbc.co.uk/partnersandsuppliers/connectedstudio/newslabs/home.html
[23] Dave Rogers, "BBC Linked Data Platform," June 5, 2013, http://www.slideshare.net/daverog/bbc-linked-data-platform.
[24] Joachim Neubert and Klaus Tochtermann, "Linked Library Data: Offering a Backbone for the Semantic Web," in *Knowledge Technology*, ed. Dickson Lukose, Abdul Rahim Ahmad, and Azizah Suliman, Communications in Computer and Information Science 295 (Springer Berlin Heidelberg, 2012), 37–45, http://link.springer.com/chapter/10.1007/978-3-642-32826-8_4.
[25] http://www.dnb.de/lds
[26] http://orcid.org/
[27] http://datahub.io/dataset?q=thesaurus
[28] AG KIM Gruppe Titeldaten Dini, "Empfehlungen Zur RDF-Repräsentation Bibliografischer Daten" (October 2, 2013), http://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=40318.
[29] http://www.w3.org/community/schemabibex/
[30] http://data.nature.com/query
[31] Dave Rogers, "The Enduring Myth of the SPARQL Endpoint," *Dave's Blog*, June 4, 2013, http://daverog.wordpress.com/2013/06/04/the-enduring-myth-of-the-sparql-endpoint/.
[32] http://www.culturegraph.org
[33] Magnus Pfeffer, "Eine Infrastruktur Für „Big Metadata"," *B.i.t. Online* 16, no. 1 (2013): 11–16.
[34] DNB, *Bericht Informationstechnik an Die AGV*, November 13, 2013, http://www.dnb.de/SharedDocs/Downloads/DE/DNB/wir/berichtDnb65.pdf.
[35] http://api.lobid.org
[36] http://zbw.eu/beta/econ-ws
[37] Joachim Neubert, "Linked Data Based Library Web Services For Economics," in *DC-2012, Kuching, Sarawak, Malaysia*, 2012, http://dcevents.dublincore.org/IntConf/dc-2012/paper/view/100.
[38] http://bibframe.org/vocab/
[39] Library of Congress, "Bibliographic Framework Initiative Update Forum - November 22, 2013," November 22, 2013, http://www.loc.gov/bibframe/media/updateforum-nov22-2013.html.
[40] Julia Hauser, Reinhold Heuvelmann, and Lars G. Svensson, "BIBFRAME: Libraries Can Lead Linked Data" (presented at the SWIB13, Hamburg, Germany, November 26, 2013), http://swib.org/swib13/slides/hauser_swib13_103.pdf.
[41] Library of Congress, "Bibliographic Framework Initiative Update Forum - November 22, 2013."

[42] http://demo.cubicweb.org/opencatfresnes; erweiterte Funktionalität und Import eigener Bibliotheken nach Registrierung: https://demo.cubicweb.org/library
[43] Agnès Simon, "The 'OpenCat' Prototype: Linking Public Libraries to National Datasets" (presented at the SWIB13, Hamburg, November 27, 2013), http://swib.org/swib13/slides/simon_swib13_104.pdf.
[44] Martin Malmsten, "Making a Library Catalogue Part of the Semantic Web," in *Proc. Int'l Conf. on Dublin Core and Metadata Applications*, 2008, http://edoc.hu-berlin.de/conferences/dc-2008/malmsten-martin-146/PDF/malmsten.pdf.
[45] http://librisbloggen.kb.se/2011/12/20/
[46] https://github.com/libris/librisxl
[47] Martin Malmsten, "Cataloguing in the Open - the Disintegration and Distribution of the Record," *JLIS.it* 4, no. 1 (January 15, 2013): 417, doi:10.4403/jlis.it-5512.
[48] http://kat.libris.kb.se
[49] http://librisbloggen.kb.se/2013/12/18/