Borst, Timo; Loehden, Aenne; Pohl, Adrian

Article

SWIB 14 - Semantic Web in Bibliotheken

ABI Technik

Suggested Citation: Borst, Timo; Loehden, Aenne; Pohl, Adrian (2015) : SWIB 14 - Semantic Web in Bibliotheken, ABI Technik, ISSN 2191-4664, De Gruyter, Berlin, Vol. 35, Iss. 1, pp. 48–52, https://doi.org/10.1515/abitech-2015-0013

Tagungsbericht

Timo Borst, Aenne Löhden und Adrian Pohl

SWIB14 – Semantic Web in Bibliotheken

DOI 10.1515/abitech-2015-0013

Zum sechsten Mal lud die internationale Fachtagung Semantic Web in Bibliotheken (SWIB) zum Austausch über Entwicklungen im Umfeld von Linked Open Data (LOD) in der Bibliothekswelt ein und brachte Vertreter von Bibliotheken, Museen, Archiven, Verbundzentralen, Forschungsinstituten, Projekten, Unternehmen und Verlagen aus 24 Ländern aus drei Kontinenten zusammen. Ausgerichtet wurde die SWIB14 wie bisher vom Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz) und der Deutschen Zentralbibliothek für Wirtschaftswissenschaften – Leibniz-Informationszentrum Wirtschaft (ZBW). Die SWIB14 fand vom 1.–3. Dezember 2014 in Bonn statt, die Räumlichkeiten wurden von der Bibliothek der Friedrich-Ebert-Stiftung zur Verfügung gestellt. Das Programm umfasste Workshops, Keynotes, Vorträge und Lightning Talks. Programm, Videomitschnitte und Folien sind über die Website[1] verfügbar, Tweets bei Twitter[2] nachvollziehbar.

1 Entitätenbildung macht aus Information durchsuchbare, aggregierbare, wiederverwendbare Daten

Bibliotheken erleichtern und strukturieren die Erschließung und Recherche von Medien schon seit einiger Zeit durch die Erfassung und Verlinkung von Datensätzen z. B. zu Titeln, Autoren und Schlagwörtern. Mit dem Aufkommen von Functional Requirements for Bibliographic Records (FRBR) kommt auch die Gliederung in Werke, Expressionen (z. B. Übersetzungen), Manifestationen (etwa bestimmte Ausgaben) und Exemplare (ein konkretes physisches Buch) hinzu. Bibliotheken erfassen und verknüpfen also bereits Entitäten. Damit lassen sie sich als Vorreiter für kommerzielle Medienvertriebe wie Amazon und Suchmaschinen wie Google ansehen, die ihre Suchergebnisse inzwischen ebenfalls nicht mehr nur als Zeichenketten betrachten, sondern zu Entitäten zusammenfassen, wie in mehreren Vorträgen der Konferenz erwähnt. Linked-Data-Technologien (LD-Technologien) helfen nun bei der Identifikation von Entitäten und der Veröffentlichung ihrer Beschreibungen. Bei Linked Data,

Abb. 1: Teilnehmer aus 24 Ländern aus drei Kontinenten bei der SWIB14 in der Friedrich-Ebert-Stiftung (Bonn) (Foto: Philippe Ramakers, Intuitive Fotografie)

1 http://swib.org/swib14/programme.php.
2 https://twitter.com/search?q=%23swib14.

Abb. 2: Internationaler Austausch der (bibliothekarischen) LOD-Community – SWIB14-Pausengespräche (Foto: Philippe Ramakers, Intuitive Fotografie)

so befand etwa Richard Wallis (OCLC), gehe es im Wesentlichen um die Auszeichnung von Dingen oder Entitäten und ihrer Zwischenverbindungen im Web der Dinge sowie die Verlinkung mit zentralen Normdaten.

Der Normdatendienst Entity Facts[3] der Deutschen Nationalbibliothek (DNB), den Christoph Böhme (DNB) vorstellte, fokussiert Entitäten. Er liefert je Entität Metadaten, die bereits zur Nuzerpräsentation aufbereitet sind. Bisher deckt der Dienst, den die Deutsche Digitale Bibliothek (DDB) verwendet, Personen ab, später sollen auch Körperschaften und Orte hinzukommen. Die Normdaten der DNB werden dabei aus externen Quellen (z. B. VIAF, IMDb, Wikipedia) angereichert. Die universelle bibliografische Datenbank WorldCat, die Kataloge von OCLC-Mitgliederbibliotheken umfasst, enthält seit 2014 auch Werke als Entitäten.

Die British Broadcasting Corporation (BBC) produziert täglich eine Fülle von Inhalten, die über verschiedene Publikationsformen und Präsentationskanäle veröffentlicht werden. Diese Inhalte seien früher isoliert erstellt und vorgehalten worden, mit in die Inhalte (z. B. HTML-Seiten) eingebundenen Themen, so Tom Grahame (BBC) – dabei bezögen sie sich oftmals auf gleiche Sachverhalte. Durch die Tool-gestützte Analyse von Inhalten und Erkennung und Disambiguierung von Entitäten wie etwa Ereignissen oder Personen (Named Entity Recognition, NER) werden diese Dinge nun identifiziert und die Inhalte erschlossen – sie lassen sich gezielt suchen und automatisch aggregieren. Inhalte zu Themen, für die sich Nutzer interessieren, wie eine Sportmannschaft oder ein Musikwettbewerb, können so automatisiert zusammengebracht und dargeboten werden.

Das Lebenswerk der Choreographin und Ballettdirektorin Pina Bausch wird in einem digitalen, LD-basierten Archiv zugänglich gemacht, berichtete Kerstin Diwisch (Hochschule Darmstadt). FRBR ordnet hier den Tanzstücken (Werk) ihre Aufführungen (Expression), Aufführungsmaterial wie Filmaufnahmen (Manifestation), und konkrete Ressourcen wie ein Videoband (Exemplar) zu – so ist alles, was zu einem bestimmten Tanzstück gehört, gebündelt und auffindbar. Um möglichst homogene Daten zu erreichen, hält ein Dublin Core Application Profile (DCAP) Vorgaben formalisiert fest.

Das Werkzeug ALIADA,[4] präsentiert von Ádám Horváth (Museum of Fine Arts Budapest), detektiert im Rahmen der automatischen Generierung von Bibliotheks- und Museums-Normdaten als LOD auch Entitäten. Und zwar erkennt es Personen, Körperschaften und Themen und strukturiert Medien gemäß FRBR. Die von Jens Mittelbach (SLUB Dresden) vorgestellte Plattform D:SWARM

3 http://www.dnb.de/entityfacts.

4 http://aliada-project.eu/.

erlaubt außer der Anreicherung, Normalisierung, Deduplizierung und Modellierung bibliothekarischer Daten auch ihre Strukturierung nach FRBR.

Einen Mangel an Entitätenbezug sieht Pascal-Nicolas Becker (Technische Universität Berlin) bei Publikationsplattformen, die digitale Objekte mit ihren Metadaten vorhalten und veröffentlichen. So speicherten viele Repositorien zu Autoren nur die Namen als Zeichenketten, anstatt durch Verknüpfung mit zentralen Normdaten die Autoren zu identifizieren, d. h. auf konkrete Personen abzubilden. Überhaupt böten Repositorien ihre Metadaten hauptsächlich wortorientiert und menschenlesbar an, während sie maschinenlesbar hinter nicht-generischen Schnittstellen versteckt seien oder gar nicht veröffentlicht würden. Dabei ließen sich ihre Metadaten sehr gut in Linked Data konvertieren und publizieren, enthielten doch Repositorien bereits strukturierte Metadaten, generierten stabile URIs (Uniform Resource Identifier) für ihre Objekte und änderten sich selten.

2 Linked-Data-Technologien helfen beim Umgang mit Divergenzen und Änderungen

Ändern sich Begriffssysteme (Knowledge Organisation System, KOS) im Laufe der Zeit, so wirkt sich dies z. B. auf die Indexierung von Medien mit den KOS oder die Abbildung auf andere KOS aus. Mit der zunehmenden Verflechtung von Vokabularen im Web der Daten scheint die genaue Kenntnis der Änderungen noch relevanter zu werden. KOS werden zwar bereits versioniert, Joachim Neubert (ZBW – Leibniz Informationszentrum Wirtschaft) vermisst aber Standards zur präzisen Kommunikation und maschinellen Verwertbarkeit von KOS-Änderungen und verweist auf das Projekt skos-history[5]. Eine Fallstudie mit dem Standard-Thesaurus Wirtschaft (STW) zeigt Umfang und Typen von Änderungen und demonstriert eine Methode, mithilfe von LD-Technologien Änderungen wie z. B. das Löschen von Deskriptoren zu dokumentieren. Wie viele andere KOS auch wird der STW als LOD im Format Simple Knowledge Organisation System (SKOS) publiziert, die relative Ähnlichkeit von SKOS-Dateien vereinfacht hier das Verfahren.

Die mehr als 280 Spracheditionen von Wikipedia weisen in Stil, geografischer Abdeckung, Qualität und Größe große Unterschiede auf und zwischen Artikeln bestehen Widersprüche, referierte Markus Krötzsch (Technische Universität Dresden). Auch bei DBpedia, dessen Daten aus Wikipedia-Artikeln automatisch extrahiert werden, gibt es je Sprache einen eigenen Datensatz. Wikidata[6] hingegen, bei dem Daten und Vokabular manuell editiert werden, besteht aus einem sprachübergreifenden Datensatz, das Vokabular liegt in 285 Sprachen vor. Im Zentrum von Wikidata stehen dabei qualifizierte Aussagen, z. B. lassen sich Personenbeziehungen mit Zeitangaben qualifizieren. So können variante Inhalte explizit nebeneinander bestehen. Auch beim Tool D:SWARM lässt sich jedes Datenelement mit zusätzlicher Information, z. B. Quelle oder Gültigkeitszeit einer Beziehung zwischen Entitäten, ausstatten. Verschiedene Quellen und heterogene Daten lassen sich so zusammenbringen.

3 Linked Open Data erhöht die Sichtbarkeit von Bibliotheken im Web

Zwar seien bei einer Suche nach Büchern via Web-Suchmaschinen bereits jetzt Bibliotheksmedien aufspürbar, jedoch nur indirekt, so Richard Wallis (OCLC). Unmittelbare Sichtbarkeit im Web erreichten Bibliotheken bzw. Bibliotheksinhalte mittels Linked Data im Format Schema.org, das für Suchmaschinen wie Google, Bing, Yahoo!, Yandex usw. lesbar sei. Für die detailliertere Beschreibung der speziellen Bibliotheksinhalte wären dann auch entsprechend qualifiziertere Vokabulare geeignet. Auch Eric Miller (Zepheira) sieht Linked Data als Mittel, die Sichtbarkeit von Bibliotheken zu erhöhen. BIBFRAME nutze, so Miller, existierende Webstandards, um (Bibliotheks-)Inhalte einheitlich im Web zu präsentieren. Zugangspunkte zu bibliothekarischen Ressourcen via Suchmaschinen seien mittlerweile durch die Publikation von bibliografischen Daten, Bestandsdaten und organisatorischen Daten möglich, stellte Dan Scott (Laurentian University) fest. Geeignete Mittel seien hier außer dem für Suchmaschinen verständlichen Vokabular Schema.org die persistente Identifikation der Ressourcen (URIs) und die Lenkung der Suchmaschinen-Spider (Sitemaps, robots.txt-Dateien). Die zunächst abgeschottete Entwicklung von Schema.org sei inzwischen offener, auch fülle die W3C Schema.org Bibliographic Extension Community Group Lücken und dokumentiere Best Practices. Viele Bibliothekssysteme und Discoverylösungen wie z. B. vuFind

5 https://github.com/jneubert/skos-history.

6 www.wikidata.org.

Abb. 3: Gelegenheit zur Diskussion bietet die SWIB auch jeweils nach den Vorträgen (Foto: Philippe Ramakers, Intuitive Fotografie)

stellen inzwischen – nicht zuletzt wegen der Programmierarbeit Scotts – automatisch Schema.org-Daten bereit.

4 Linked-Data-basierte Werkzeuge und Dienste erleichtern die Arbeit im Kultur(erbe)sektor und senken Technik- und Formatbarrieren

Neue Dienste und Werkzeuge helfen Bibliothekaren und anderen Kulturarbeitern bei ihren täglichen Aufgaben – sei es beim Umgang mit Normdaten, Thesauri oder Artikeln, bei Generierung, Recherche oder Publikation – ganz im Sinne der Keynote von Dorothea Salo (University of Wisconsin-Madison) bei der SWIB13.

Osma Suominen (Finnische Nationalbibliothek) stellte den freien, Web-basierten SKOS-Browser Skosmos[7] vor. Skosmos veröffentlicht kontrollierte Vokabulare wie Thesauri und Normdaten, die Nutzerschnittstelle unterstützt Stöbern und Suchen und visualisiert die Konzepthierarchie, eine Entwicklerschnittstelle liefert Linked Data in verschiedenen Formaten. Skosmos unterstützt also konkret die Arbeit mit kontrollierten Vokabularen, wie die Verschlagwortung, die Pflege von Vokabularen oder das Mapping zwischen Vokabularen, aber auch die Einbindung z. B. in Annotationssysteme. Der Thesaurus Global Agricultural Concept Scheme (GACS) wird beispielsweise mit Skosmos präsentiert werden, der Thesaurus AGROVOC[8] ist es bereits.

Überschneidungen zwischen kontrollierten Vokabularen sind nicht ungewöhnlich. Thesauri mit untereinander dubletten Termen seien jedoch ineffizient zu pflegen und mit ihnen erschlossene Inhalte ineffizient zu durchsuchen, konstatierte Suominen. Eine Machbarkeitsstudie unter Verwendung von LD-Technologien untersuchte die virtuelle Zusammenführung dreier stark überlappender landwirtschaftlicher Thesauri – AGROVOC, CAB Thesaurus und NAL Thesaurus. Entwickelt wurde ein halbautomatisches Mapping-Verfahren, auch wurden Werkzeuge zur Erkennung und Bereinigung von Unsauberkeiten in SKOS-Dateien, wie z. B. Zyklen in Hierarchien oder fehlende Labels, eingebunden. Der dabei entstehende Thesaurus GACS umfasst im Kern die am meisten verwendeten 10 000 Terme der drei Ausgangsvokabulare. Aus validen Mappings zwischen den drei Thesauri werden dabei neue Konzepte erzeugt. GACS kann AGROVOC, CAB Thesaurus und NAL Thesaurus nicht ersetzen, verbindet sie aber. Auch könnte GACS sie gegenseitig mit mehrsprachigen Labels bereichern, da ihre Sprachabdeckung sehr differiert.

7 https://github.com/NatLibFi/Skosmos.

8 http://aims.fao.org/standards/agrovoc/functionalities/search.

Der bereits erwähnte DNB-Dienst Entity Facts liefert seine Daten so, dass sie sowohl aus Technik- als auch aus Formatsicht intuitiv nachnutzbar sind und leicht in Nutzerpräsentationen eingebunden werden können. So sind die Daten maschinenlesbar und mehrsprachig über eine einfache JSON-Schnittstelle abrufbar. Bibliothekarische Ansetzungen z. B. bei Namen werden in eine nutzerlesbare Form umgesetzt, Zeitangaben in lokaler Schreibweise ausgegeben, für Nutzer irrelevante Metadaten ausgefiltert.

Für eine Linked-Data-Plattform zu historischen dänischen Zeitungen wurden drei gedruckte Bände mit Metadaten gescannt, referierte Martynas Jusevicius (UAB Linked Data). Aus den Scans wurden mithilfe optischer Zeichenerkennung (OCR) und XSLT-Transformation automatisiert Metadaten in LD-Formaten ermittelt. Dies ermöglicht z. B. facettiertes Suchen in den Zeitungen und die Darstellung in interaktiven Landkarten. Das dabei eingesetzte Tool Graphity[9] unterstützt bei der Entwicklung nutzerfreundlicher LD-Applikationen. Die freien Datenmanagementtools ALIADA und D:SWARM sind dediziert auf Fachkräfte statt auf Techniker ausgerichtet und ohne Programmierkenntnisse bedienbar.

Das Linked-Data-basierte und Entitäten-fokussierte Management von Inhalten der BBC gestattet es den Journalisten, Ressourcen gezielt zu recherchieren und automatisch in Präsentationskanäle einbinden zu lassen.

5 Das Semantic Web ändert Möglichkeiten und Rollen von Nutzern

Das Semantic Web erweitert einerseits die Möglichkeiten von Nutzerrecherchen. So lässt sich durch verknüpfte Thesauri bestandsübergreifend suchen, und die Thesauri gewinnen durch generische, LD-basierte Viewer wie Skosmos an Verständlichkeit. Andererseits entstehen auf Basis von LD-Technologien Umgebungen, die Nutzer an der Erschließung von Inhalten und an der Pflege von Metadaten beteiligen.

Wikipedia eigne sich vor allem zum Lesen von Texten durch Menschen, Informationen zu extrahieren sei eher schwierig, so Krötzsch. Als strukturiertere Zugänge bietet Wikipedia Infoboxen, Seiten zur Begriffsklärung und Listen. Die vielen Listen beantworten eine große Anzahl von Fragen, aber eben nur vorgegebene. Wikidata dagegen soll frei befragbare Linked Data bereitstellen. Die Daten werden einerseits von den meisten anderen Wikimedia-Projekten (z. B. Wikivoyage, Wikiquote) genutzt, sind andererseits aber auch sofort Benutzern dienlich. Denn mit ihrer Darstellbarkeit in allen 285 Wikidata-Sprachen stehen, falls zu einem Thema kein Material in der gewünschten Sprache zur Hand ist, doch zumindest Daten zur Verfügung. Aber auch aktiv setzen sich Nutzer hier mit Daten auseinander, indem sie manuell Daten und Vokabular editieren.

Cristina Sarasua (Universität Koblenz-Landau) befand, dass im Kulturerbe-Bereich mit seiner inhärenten Vielsprachigkeit, differierenden Namenskonventionen und vielen Fachgebieten Nutzer mit Gewinn kleine Aufgaben zur Anreicherung der Datenbasis übernehmen könnten. „Microtask Crowdsourcing“, wie sie es nannte, habe großes Potential. Ob sich eine konkrete Aufgabenstellung dafür eigne, hänge vor allem von den Daten ab.

Inhalte und Annotationen wie Kommentare oder Rezensionen seien im Web oft getrennt (Twitter, Facebook) oder intern proprietär gespeichert (ePUB, Kindle), stellte Dan Whaley (Hypothes.is) fest. Er verwies auf verschiedene Ansätze, Annotationen im Web zu standardisieren. Hypothes.is ist eine offene Plattform zur Annotation von Webinhalten bis auf Satz- und Wortebene. Als RDF-Vokabular kommt das Open Annotation Data Model der W3C Web Annotation Working Group zum Einsatz, das auch beim W3C die Annotation von Spezifikationen ermöglichen soll. Digitale Annotationen sind selbst adressierbare Webinhalte, sie können also verknüpft, gesucht, rückverfolgt oder annotiert werden.

Formate, Werkzeuge und Dienste des Semantic Web bringen Bibliotheken und ihre Nutzer voran – wie die Konferenz erneut zeigte. Die Folgeveranstaltung SWIB15 wird vom 23.–25. November 2015 in Hamburg stattfinden.

Dr. Timo Borst
Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW)
Düsternbrooker Weg 120
24105 Kiel
t.borst@zbw.eu

Aenne Löhden
Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW)
Düsternbrooker Weg 120
24105 Kiel
a.loehden@zbw.eu

Adrian Pohl
Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz)
Jülicher Straße 6
50674 Köln
pohl@hbz-nrw.de

9 http://graphityhq.com.