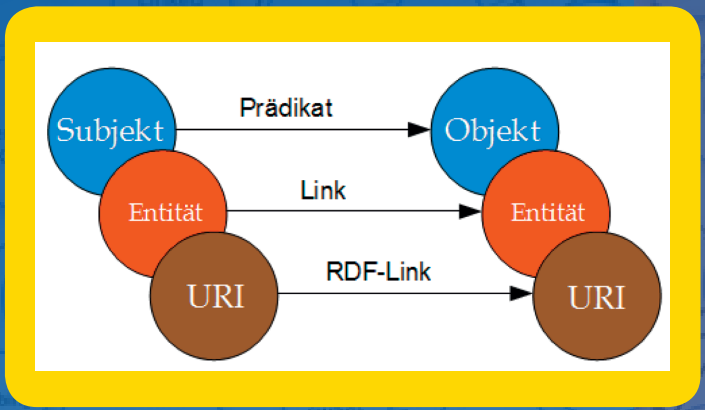
BAND 33

B.I.T.online INNOVATIV

Linked Open Library Data

Bibliographische Daten und ihre Zugänglichkeit im Web der Daten



INNOVATIONSPREIS 2011

Fabian M. Fürste

B.I.T.online – Innovativ

Band 33

B.I.T.online – Innovativ

Herausgegeben

von

Rolf Fuhlrott

Ute Krauß-Leichert

Christoph-Hubert Schütte

Band 33

Innovationspreis 2011

Linked Open Library Data
Bibliographische Daten
und ihre Zugänglichkeit im Web der Daten

2011

Verlag: Dinges & Frick GmbH, Wiesbaden

Innovationspreis 2011

Linked Open Library Data

Bibliographische Daten und ihre Zugänglichkeit im Web der Daten

von

FABIAN M. FÜRSTE

2011

Verlag: Dinges & Frick GmbH, Wiesbaden

B.I.T.online – Innovativ

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.
ISBN 978-3-934997-36-3

ISBN 978-3-934997-36-3
ISSN 1615-1577

Satz und Druck: Dinges & Frick GmbH, Wiesbaden
Printed in Germany

Vorwort

B.I.T.online Innovationspreis 2011

Themen der (nahen) Zukunft:
Freier, direkter Onlinezugriff auf Katalogdaten, Bibliotheksapplikationen für Smartphones und notwendiges Krisenmanagement

Die Preisträger des B.I.T. online Innovationspreises 2011 sind in diesem Jahr – für unser Berufsumfeld eher untypisch – drei männliche Absolventen bibliothekarischer Studiengänge. Repräsentativ dagegen für die im Umbruch befindliche Studienlandschaft ist die sich unbeabsichtigt ergebende Auszeichnung je einer Bachelor-, Diplom- und Magisterarbeit.
Ausgezeichnet wurden in alphabetischer Reihenfolge:

Drechsler, Ralf: Krisen-PR für Bibliotheken in finanziellen Notlagen: Handlungsempfehlungen für die Krisenkommunikation Öffentlicher Bibliotheken
(Bibliotheks-und Informationsmanagement Department Information, Hochschule für angewandte Wissenschaften Hamburg)

Fürste, Fabian: Linked Open Library Data. Bibliographische Daten und ihre Zugänglichkeit im Web der Daten
(Institut für Bibliotheks-und Informationswissenschaft, Humboldt-Universität zu Berlin)

Pohla, Hans-Bodo: Untersuchung bibliothekarischer Applikationen für Mobiltelefone hinsichtlich der technischen Realisierung und des Nutzens
(Fakultät für Informations- und Kommunikationswissenschaften, Fachhochschule Köln)

Auch die Themen der Abschlussarbeiten sind kennzeichnend für eine in vielfältiger Hinsicht im Umbruch befindliche Bibliothekslandschaft.1

Öffentliche Bibliotheken, ungeachtet einzelner Bibliotheksgesetze der Länder, als nach wie vor freiwillige Leistung der Kommunen geraten aufgrund schwieriger finanzieller Rahmenbedingungen der Städte und Gemeinden vermehrt in finanzielle Nöte bis hin zu beabsichtigten Schließungsmaßnahmen. Drohende Krisen, frühzeitig zu erkennen an allgemeiner Zuweisungsreduzierung, Kürzungen von Öffnungszeiten, Stelleneinsparungen, aber ggf. auch rückläufigen Besucherzahlen oder Entleihungen, sind im Zunehmen begriffen.

1 Der Verlag veröffentlicht die drei Arbeiten in seiner Buchreihe B.I.T.online INNOVATIV 2011 als Bd. 32 (Drechsler), Bd. 33 (Fürste) und Bd. 34 (Pohla).

Welche Möglichkeiten der Krisenkommunikation und des Krisenmanagements betroffenen Stadtbibliotheken offenstehen, dem geht die Bachelorarbeit von ***Ralf Drechsler*** auch unter Heranziehung von Fallbeispielen erfolgreicher Krisen-PR aus der deutschen Unternehmenslandschaft (von TUI bis Humana) und Interviews mit Experten aus der Krisenforschung und Bibliotheken nach. Eingebettet in die bürokratisch-hierarchische Organisation öffentlicher Verwaltung unterliegen Öffentliche Bibliotheken im Gegensatz zur freien Wirtschaft starken Kommunikationseinschränkungen, können ohne Loyalitätsverletzung gegenüber dem Träger nur durch Multiplikatoren der eigenen Botschaft in lobbyistischen Zielgruppen die Bevölkerung überzeugen und aktivieren, gegen Kürzungen bzw. für den Bibliothekserhalt einzustehen. Das Gelingen setzt jedoch voraus, dass im Vorbild möglicher Krisen im Rahmen kontinuierlicher, möglichst hauptamtlicher PR-und Öffentlichkeitsarbeit notwendige Kontakte zu einflussreichen Personen der Städte aufgebaut und gepflegt wurden sowie Aufgaben und Notwendigkeit der Bibliothek in Presse und Öffentlichkeit verankert wurden.

Wenn Interessierte heute **bibliographische Daten** über das Internet suchen, so sind in Bezug auf die Bibliothek einzig deren Kataloge selbst auffindbar, die eigentlichen Daten bleiben im Deep Web verborgen und werden über die gängige Suchmaschinen nicht indexiert. Diese indirekte Bereitstellung erschwert nicht nur die Auffindbarkeit, sondern wird auch dem Wert der durchweg akribisch erstellten und hochqualitativen Daten nicht gerecht. Auch bleiben die Online-Bibliothekskataloge ungeachtet der Anreicherung mit Rezensionen und Titelbildern auch externer Anbieter sowie vereinzelter Annotationen durch Nutzer in erster Linie Bestandsanzeiger, an die lediglich über definierte Benutzerschnittstellen Suchanfragen an den Server gestellt werden können. Metadaten stehen nur in der vordefinierten Sicht eines Katalogeintrages zur Verfügung, die Weiterverwendung in anderen Umgebungen ist nicht sichergestellt und unter Umständen auch nicht gewünscht – Schnittstellen wie beispielsweise Z3950 wirken als technische Barriere.
Aufbauend auf dem geschilderten Status Quo wird in der Berliner Magisterarbeit von ***Fabian Fürste*** dargestellt, wie Linked Open Data als Alternative eines nahtlosen Trägermodells unter Harmonisierung der Vielzahl mittlerweise entstandener Formatstandards und ihren implizierten Datenmodellen (MARC, METS, Dublin Core…) geeignet wäre, die Bedürfnisse einer großen Nutzerschaft zu bedienen.
Die Möglichkeit, bibliographische Daten in einem gemeinsamen Datenmodell miteinander in beliebige Beziehungen setzen zu können, bietet die notwendigen Voraussetzungen, bisherige Schranken externer Datenkommunikation abzutragen, die Indexierung und Verarbeitung bibliographischer Daten durch Suchmaschinen zu ermöglichen. Eine Adaptierung bietet ferner eine Reihe weiterer nützlicher Nebeneffekte, so

können bibliographische Daten nach völlig neuen Kriterien durchsucht werden, als Beispiel nennt Fürste z.B. die Suche nach literarischen Erstlingswerken von Autoren, die nicht älter als 40 Jahre alt wurden. Unter positiver Resonanz der Fachöffentlichkeit haben bereits einige bibliothekarische Einrichtungen den Weg von Open Data beschritten: die Universitätsbibliothek Huddersfield (USA), die Bibliothek des Genfer Forschungszentrums CERN und im März 2010 die Universitäts-und Stadtbibliothek Köln.

Ob und wie Bibliotheken ihre zunehmend den Smartphone-Nutzern zuzurechnenden potentiellen Kunden mittels bibliothekarischer Applikationen für Mobiltelefone erreichen können, steht im Focus von ***Hans-Bodo Pohlas*** an der Fachhochschule Köln angefertigter Diplomarbeit. Insgesamt 27 der derzeit für den Nutzer kostenfrei angebotenen Applikationen überwiegend amerikanischer Bibliotheken wurden untersucht. Hitliste waren Zugriffe auf Nutzerkonto und den Bibliothekskatalog, aber auch auf Empfehlungsdienste, Datenbanken und Nachschlagewerde – wünschenswert des weiteren die Entwicklungen von Lokalisierungsangeboten der nächstgelegenen Bibliotheken und die Zielführung ggf. bei entsprechender Visualisierung der Öffnungszeiten in der Kartierung, noch steigerungsfähig bis zur Hinführung vom Katalogeintrag zur richtigen Etage, Raum und Regal der Bibliothek.
Im Vergleich von Applikationen und der mobilen Webseite sind zu Gunsten ersterer wenig Vorteile zu erkennen, höchstens in Bezug auf die Kontaktaufnahme. Ein Handicap ist in beiden Fällen nicht mehr mangelnde Internetgeschwindigkeit oder Speichermöglichkeiten, sondern die Displaygröße beispielsweise bei Kataloganfragen. Noch gestaltet sich infolge der großen Heterogenität der Systeme und unterschiedlichen Entwicklungsmöglichkeiten die technische Realisierung problematisch, da sie nur Sinn macht, wenn eine breite Masse von Smartphone-Nutzern erreicht werden kann. Grundsätzlich könnte aber in naher Zukunft jede Bibliothek bei überschaubarem Aufwand auf eine auf ihre Bedürfnisse zugeschnittene Applikation verfügen und diese würde dann ganz selbstverständlich ins Portfolio gehören.

Karin Holste-Flinspach
Kommission für Ausbildung und Berufsbilder (Vorsitzende)

Linked Open Library Data

Bibliographische Daten und ihre Zugänglichkeit im Web der Daten

Magisterarbeit

Humboldt-Universität zu Berlin
Philosophische Fakultät I
Institut für Bibliotheks- und Informationswissenschaft

von

FABIAN M. FÜRSTE

Gutachter/in: 1. Prof. Dr. Stefan Gradmann
2. Dr. Uwe Müller
Datum der Einreichung: 25. April 2010

Für Kat und Cosmo.

Inhaltsverzeichnis

1 Einleitung

Die Welt der Bibliotheken sah sich in den vergangenen Jahren mit einer Vielzahl an Problemen konfrontiert, die in ihrer Gesamtheit an der Selbstverortung der bibliothekarischen Zunft genagt haben. Bibliotheken galten lange als die akademischen Leitinstitutionen ihrer Zeit[1], die mit ihrer Wissensware einen großen Teil des Informationsbedarfs deckten. Diese Zeiten sind vorbei. Zurechtgestutzt auf die Rolle von reinen Informationsdienstleistern sehen sich Bibliotheken mittlerweile einer Vielzahl privatwirtschaftlicher Konkurrenz auf dem Feld der Informationsbeschaffung ausgesetzt. Dieser hinsichtlich Wirtschaftskraft und Reichweite ungleiche Kampf wird auf absehbare Zeit nicht zu gewinnen sein. Doch nicht nur die reine Informationssuche ist zu großen Teilen in die populären Suchportale abgewandert. Auch haben spezialisierte und bibliotheksfremde Anbieter vor allem in den Domänen der Natur- und Rechtswissenschaften einen großen Teil der qualitativen Informationsbeschaffung übernommen.

Dies hat zu einer tiefen Verunsicherung geführt, die seit einigen Jahren die Diskussion bestimmt. Die scheinbar grenzenlose Euphorie gegenüber neuen Impulsen wie „Bibliothek 2.0“ lässt erkennen, wie sehr sich Bibliotheken nach einem rettenden Strohhalm in der unübersichtlichen digitalen Informationswelt sehnen. „Bibliothek 2.0“ entwickelte sich im Rahmen des „Web 2.0“-Hype der 2000er Jahre, im Rahmen dessen sich das WWW von einem passiven Informationskanal zu einem aktiven Kommunikationskanal transformierte. Dabei wurden unter Verwendung bestehender und standardisierter Webtechnologie (vor allem JavaScript und XML als AJaX und HTTP-basierte Webservices) interaktive Feedback-Mechanismen geschaffen, die zu einer fulminanten Erweiterung der webbasierten Kommunikations- und Kollaborationsmöglichkeiten führten und durch Weblogs, Podcasts, Wikis etc. das sogenannte „soziale Web“ entstehen ließen. Ein zweites wesentliches Prinzip war der Mashup-Gedanke, die Mixtur von Daten aus unterschiedlichen Domänen über vordefinierte Schnittstellen (APIs).

In der Hoffnung, an diesen Entwicklungen zu partizipieren, wurden in der Folge viele Bibliothekskataloge mit Rezensionen und Titelbildern externer und privatwirtschaftlicher Anbieter wie Amazon angereichert. Auch die Annotation von bibliographischen Ressourcen durch NutzerInnen (Tagging) wurde

[1] Vgl. [Gradmann 2005], 99f

begeistert aufgenommen, blieb jedoch ohne nennenswerte Reichweite, da der institutionelle Bibliothekskatalog in erster Linie noch immer als Bestandsanzeiger konzipiert und nicht wie z.B. das Portal LibraryThing als Community gedacht wird. Dabei hat im deutschen Raum das beluga-Projekt der Staats- und Universitätsbibliothek Hamburg in einer begleitenden Nutzerstudie einige interessante Erkenntnisse zutage gefördert: so wurde angedeutet, dass der Katalog *„für das Teilen von bibliografischen Informationen nicht der beste Ort"* sei, und auch die individuelle Verschlagwortung durch Tags sowie die Möglichkeit der eigenen Rezension seien *„nicht wirklich begeistert"* und *„indifferent"* aufgenommen worden. Die Anlehnung an soziale Plattformen erschloss sich den Probanden demnach nicht.[2] Als eine wichtige und zentrale Komponente wurde vor allem *„die Bedeutung von offenen, standardisierten Schnittstellen für den Export von bibliografischen Daten in andere Umgebungen"* bestätigt.[3] Unter Berufung auf Untersuchungen der dänischen Staatsbibliothek im Rahmen des Summa-Projekts wurden weitere aus Nutzersicht wünschenswerte Funktionen eines modernen Bibliothekskatalogs herausgearbeitet.[4] Wichtige Faktoren seien demnach die *„Anreicherung der bibliografischen Informationen mit nichtbibliothekarischen Inhalten"*, Visualisierungen für ein optimiertes *„Nutzungserlebnis"* und *„offene, standardkonforme Schnittstellen für den Austausch von Inhalten mit anderen Systemen"* unter Berücksichtigung *„von Mikroformaten und anderen Web-Standards"*.[5] Die dänische Studie ermittelte darüber hinaus den Wunsch an Datenbankhersteller, eine standardisierte Schnittstelle für ihre Produkte anzubieten und so einen homogenen Sucheinstieg zu ermöglichen.[6] Insgesamt wurde konstatiert, dass eine bibliothekarische Beschäftigung mit evaluierten sozialen Funktionalitäten in Katalogen – trotz geringerer Akzeptanz – eine wünschenswerte Entwicklung sei, da ein *„fast übergroßes"* und *„generelles Bedürfnis nach einer zusätzlichen Aufbereitung von bibliografischen Daten"* bestünde, vor allem *„in Form von thematischen Zusammenstellungen in speziellen Nutzungszusammenhängen"*. Notwendig sei dafür jedoch die Entwicklung von *„offenen Schnittstellen, Standards und Lizenzmodellen"* zur *„Weiterverwendung von bibliografischen Daten in anderen Umgebungen"*.[7]

---

[2] Vgl. [Christensen 2009], 535

[3] Vgl. [Christensen 2009], 533

[4] http://www.statsbiblioteket.dk/summa/

[5] Vgl. [Christensen 2009], 529

[6] Vgl. [Akselbo et al. 2006], 14

[7] Vgl. [Christensen 2009], 536

Das „Web 2.0"-Paradigma bietet den Bibliotheken vor allem Möglichkeiten, ihre Außendarstellung durch nutzerfreundlichere Gestaltung der Datenpräsentation und Berücksichtigung gängiger ästhetischer Maximen wesentlich zu verbessern. Es ist darüber hinaus ein starkes Signal für eine verstärkte Beschäftigung mit Webtechnologie. Und doch: in seiner gängigen Form bietet es den Bibliotheken keine entscheidenden Impulse für die drängenden Probleme dieser Zeit an, denn das eigentliche Kernproblem der Bibliotheken ist weniger ein Problem der Präsentation als eines der *Repräsentation* ihrer Daten.

Das World Wide Web (WWW) hat sämtliche Bereiche des Lebens erfasst und radikal verändert. Seit seiner Entstehung 1989, erfunden durch Tim Berners-Lee am Genfer CERN-Forschungszentrum, ist es rapide und exponentiell gewachsen und verdoppelt, angelehnt an das Moore'sche Gesetz, zirka alle 5 Jahre seine Größe.[8] Bibliotheken haben in diesem Informationsraum lange Zeit keine bemerkenswerte Rolle gespielt. Zwei wesentliche Gründe sind dafür maßgeblich verantwortlich: die traditionelle technologische Fundierung bibliographischer Daten und das ihnen zugrundeliegende Datenmodell, das gravierende Auswirkungen auf seine webbasierte (Nicht-)Zugänglichkeit hat.

Interoperabilität und Offenheit ihrer Metadaten wurden lange Zeit von den Bibliotheken vernachlässigt. Die traditionelle Praxis der Datenverteilung besteht im Austausch von Daten über Schnittstellen wie z39.50. Als diese interbibliothekarischen Dienste in den 2000er Jahren auf den allgemein verwendbaren Kommunikationskanal WWW portiert wurden, blieb die Ebene der Daten davon unberührt. Ihre Heterogenität resultiert aus einer Datenbehandlung, die *„fast ausschließlich auf Ebene der Eingabe und Anzeige"* stattfindet, *„während die eigentliche Datenverarbeitung in geschlossene Systeme - verbannt wird"* [9] und durch die Entwicklung unterschiedlichster Formate für verschiedenste Anwendungsszenarios stetig gewachsen ist. MARCXML, MABXML, MODS, MADS, METS, DublinCore sowie PREMIS und ONIX sind dabei nur die bekanntesten Trägerformate. Obwohl die meisten dieser Formate in einer XML-Serialisierung vorliegen, ist ihre Struktur höchst variabel. Deshalb bedarf es für eine wirkliche Interoperabilität idealerweise eines gemeinsamen Datenmodells.

---

[8] Vgl. [Zhang et al. 2008]

[9] Vgl. [Voß 2009a]

Auch eine Offenheit der Daten ist faktisch nicht gegeben. Nach einer Definition der *Open Knowledge Foundation* können Daten als offen bezeichnet werden,

- die frei (im Internet) in ihrer Gesamtheit zugänglich sind
- die eine freie Folgenutzung und Weiterverteilung zulassen
- unter der Voraussetzung, dass die genannten Schritte ohne Überwindung technischer Hindernisse erfolgen können, beispielsweise durch Verfügbarmachung in einem offenen Format [10]

Metadaten stehen weder in ihrer Gesamtheit zur Verfügung – sondern nur in singulärer Form und der vordefinierten Sicht eines Katalogeintrags – noch können sie auf beliebige Weise nachgenutzt und weiterverteilt werden. Die zur Verfügung stehenden Schnittstellen wie z39.50 und SRU/W sind technologische Barrieren und verhindern eine freie Nutzung der Daten.

Die vorliegende Arbeit beschäftigt sich mit diesem Status Quo und stellt ihm eine interoperable und offene Alternative für die Kodierung bibliographischer Information entgegen. Sie untergliedert sich in vier große Abschnitte. Nach einer kursorischen Bestandsaufnahme der Fundierung bibliographischer Daten und ihren webbasierten Zugangsmöglichkeiten werden Gründe für ihre direkte Veröffentlichung aufgeführt. Anschließend wird ein unter dem Oberbegriff *Semantic Web* in letzter Zeit populär gewordener Ansatz vorgestellt, der das Potential hat, das Datendilemma der Bibliothekswelt zu entschärfen: *Linked (Open) Data*. Seine Prinzipien werden in einem Praxisteil vorgestellt und exemplarisch auf bibliographische Daten angewandt. Die Arbeit wird abgerundet von einem Ausblick auf Konsequenzen von *Linked (Open) Data* für die bibliothekarische Arbeit.[11]

[10] Vgl. [Open Knowledge Foundation 2010]

[11] Im Rahmen dieser Arbeit bleibt der Bereich der "Digitalen Bibliotheken" aufgrund seiner unscharfen Definition von der Behandlung ausgenommen. Die vorgestellten Konzepte ließen sich jedoch leicht auf diesen Themenkomplex ausweiten.

2 Bibliographische Daten

2.1 Was sind bibliographische Daten?

Bibliotheken und andere „Gedächtnisinstitutionen" wie Archive und Museen[12] haben über Jahrzehnte hinweg ihre Objekte mit einer riesigen Menge an hochwertigen Metadaten ausgezeichnet. Dazu gehören (neben Bestands- und Nutzungsdaten)

- *Titelangaben*, die die beschreibenden Eigenschaften von bibliographischen Objekten beinhalten.
- *Normdaten und Thesauri*, die zu den kontrollierten Vokabularen gehören und nicht-hierarchische Personen- und Schlagwortverzeichnisse umfassen, in denen Ansetzungsformen für die inhaltliche Erschließung (Sacherschließung) von bibliographischen Objekten enthalten sind.
- *Klassifikationen und Taxonomien*, die ebenfalls zu den kontrollierten Vokabularen gehören und monohierarchische Begriffssysteme für die Systematik von Themenkomplexen (mit Entsprechungen in den Thesauri) umfassen.

Zu den Normdaten zählen in Deutschland die Schlagwortnormdatei SWD, die Gemeinsame Körperschaftsdatei GKD und die Personennamendatei PND. Auf internationaler Ebene sind u.a. die *Library of Congress Subject Headings* (LCSH), das französische *Répertoire d'autorité-matière encyclopédique et alphabétique unifié* (RAMEAU) und der *UNESCO Thesaurus* von Bedeutung. Speziellere Vokabulare sind z.B. die *Medical Subject Headings* (MeSH) und der *Getty Thesaurus of Geographic Names*. Wichtige taxonomische Klassifikationen bilden in Deutschland die Regensburger Verbundklassifikation und international die Klassifikation der Library of Congress (LCC), die Dewey Dezimalklassifikation (DDC) und die Universalklassifikation UDC.

2.2 Maschinelle Austauschformate

Metadaten und kontrollierte Vokabulare werden über Integrierte Bibliothekssysteme (ILS) organisiert und in geschlossenen relationalen Datenbanken vorgehalten, deren Tabellenrelationen und Datenfelder spezifische und proprietäre Strukturen und Semantiken aufweisen. Um den interoperablen Austausch

[12] Vgl. [Dempsey 2000]

mit anderen Bibliotheken zu ermöglichen, wurden im Zuge der Bibliotheksautomatisierung in den 1960er Jahren maschinelle Austauschformate entwickelt. Sie ermöglichen die Datenextraktion und -integration zwischen heterogenen Bibliothekssystemen, die für eine Verarbeitung konfiguriert sind. Die Übertragung erfolgt über physische Datenträger oder netzbasierte bibliothekarische Schnittstellen.

2.2.1 MARC

MARC (*Machine-Readable Cataloging*) gilt aufgrund seiner flächendeckenden Unterstützung durch die Softwareinfrastruktur auf internationaler Ebene als das Austauschformat in Bibliotheken. Vielfach wird es auch als internes Format eingesetzt; vor allem in den USA hat es den Status eines „Katalogformats". MARC ist feld-orientiert und sein Datenmodell baut auf dem ISO-Standard 2709 (*Format for Bibliographic Information Interchange on Magnetic Tape*) auf. MARC wurde in den 1960er Jahren an der amerikanischen Library of Congress entwickelt, um der damaligen technischen Revolution im Bibliotheksbereich – Datenaustausch über Lochkarten und Magnetbänder – Rechnung zu tragen und liegt aktuell in der Version MARC21 vor. Der ISO-Standard 2709 beschreibt den Aufbau von Austauschformaten in drei Komponenten:[13]

- *Record Structure*: die „physische Struktur" des Datencontainers für die Speicherung
- *Content Designation*: die Codekonventionen zur Identifizierung der einzelnen Datenfelder innerhalb des Datencontainers
- *Content*: der durch Regelwerke wie AACR2 oder RAK-WB definierte syntaktische Aufbau der bibliographischen Inhalte

MARC ermöglicht auf dieser Grundlage die strukturierte Repräsentation bibliographischer Daten in spezifischen Formaten, von denen uns v.a. die folgenden interessieren sollen:

- Bibliographische Daten • *Format for Bibliographic Data*
- Normdaten • *Format for Authority Data*

[13] Vgl. [Chowdury 2007], 48

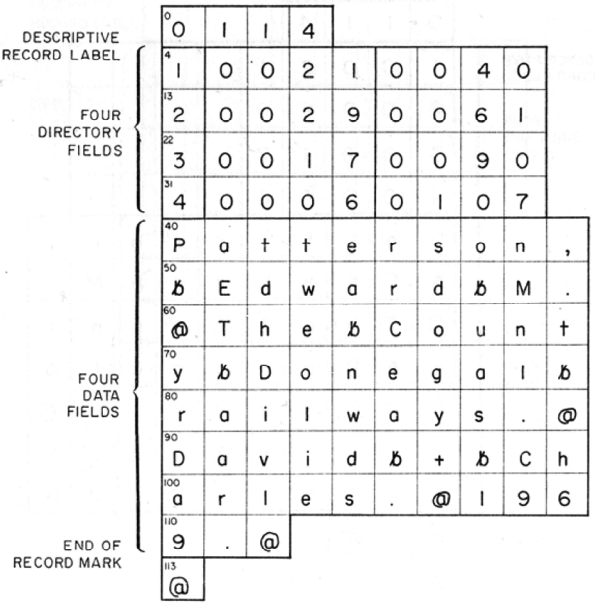


Abb. 1: Ursprüngliche Organisation eines MARC-Datensatzes [14]

Die Organisation der Daten erfolgte ursprünglich wie in [Abb. 1] dargestellt. Die gegenwärtig gängige Repräsentation eines bibliographischen MARC-Datensatzes entspricht der folgenden Struktur, wie sie über die Library of Congress verfügbar ist:

[14] Zitiert nach: [Kimber 1974], 74

```
=LDR  01063cam a22002771  4500
=001  3587391
=005  20050831151926.0
=008  711012s1966\\\\pau\\\\\\\\\\\\000\1\eng\\
=035  \\$9(DLC)   66012340
=906  \\$a7$bcbc$corignew$du$eocip$f19$gy-gencatlg
=010  \\$a   66012340
=020  \\$a0397004184
=040  \\$aDLC$cDLC$dDLC
=050  00$aPZ4.P997$bCr$aPS3566.Y55
=051  \\$aPS3566.Y55$bC79 1966$cCopy 3.
=100  1\$aPynchon, Thomas.
=245  14$aThe crying of lot 49.
=250  \\$a[1st ed.]
=260  \\$aPhiladelphia,$bLippincott$c[1966]
=300  \\$a183 p.$c21 cm.
=500  \\$a"A portion of this novel was first published in Esquire magazine
        under the title: The world (this one), the flesh (Mrs. Oedipa Maas),
        and the testament of Pierce Inverarity. Another portion has appeared
        in Cavalier."
=650  \0$aAdministration of estates$xFiction.
=650  \0$aMarried women$xFiction.
=651  \0$aCalifornia$xFiction.
=991  \\$bc-GenColl$hPZ4.P997$iCr$p0002131836A$tCopy 1$wBOOKS
=991  \\$bc-RareBook$hPS3566.Y55$iC79 1966$tCopy 1$wBOOKS
```

Abb. 2: Repräsentation von bibliographischen Daten
im MARC Format for Bibliographic Data

Erkennbar ist, dass das Format aus einer Zeit stammt, in der Speicherplatz ein hohes Gut war. Die Daten sind für die damalige maschinelle Verarbeitung effizient strukturiert und können auf dieser technologischen Basis von Bibliothekssystemen verarbeitet werden. Das Beispiel aus [Abb. 2] enthält sämtliche Metadaten einer bibliographischen Einheit: dazu zählen z.B. die Ansetzungsform des Autors (100), der Titel (245), Veröffentlichungsangaben (260), vergebene Schlagworte (650, 651) und klassifikatorische Einordnung (040-051). Der Datensatz im *Format for Authority Data* für die Ansetzungsform eines Autorennamens besitzt den gleichen ISO-Aufbau mit anderen Feldcodes:

3 Grundlagen des Semantic Web

3.1 Das World Wide Web

Informationen liegen im WWW vor allem in semi-strukturierten Dokumenten vor, die in einer Auszeichnungssprache wie (X)HTML für die Präsentation der Daten optimiert sind. Um die Relevanz eines solchen Dokuments einschätzen zu können, muss es jeweils vom menschlichen Auge rezipiert, in seiner Bedeutung erfasst, in einen Kontext eingeordnet und manuell-intellektuell einer „Weiterverarbeitung" zugeführt werden. Eine Automatisierung dieser Arbeitsschritte ist auf dieser Grundlage nicht möglich. Ein einfaches Beispiel, das wir im Folgenden verwenden und ausbauen wollen, stellt der Quellcode eines Suchergebnisses in einem Bibliothekskatalog dar, wie in [Abb. 14] dargestellt.

```
<h1 class="title">The crying of Lot 49</h1>
<div>Verfasser/in: Thomas Pynchon</div>
<div>Verlag: Philadelphia : J. B. Lippincott Company, 1966.</div>
<div>ISBN: 0-397-00418-4</div>
```

Abb. 14: HTML-Repräsentation bibliographischer Daten

Die Gesamtinformation ist für einen Computer nicht erfassbar, da die einzelnen Fakten in ihrer Identität und ihrer Bedeutung nicht zusammengezogen und gemeinsam interpretiert werden können. Was ist hier aber mit Bedeutung gemeint?

Das Web besteht grundsätzlich aus zwei Komponenten: Entitäten und Graphen. Entitäten sind im Web-Kontext sämtliche Dateien, die über einen eindeutigen Identifier adressiert werden können. Im Web-Kontext können das u.a. sein:

HTML	http://example.org/index.html
PDF	http://example.org/vita.pdf
Multimedia	http://example.org/interview.mp3
XML	http://example.org/data.xml

Verschiedene Entitäten werden im WWW durch Hyperlinks verbunden. Die Verbindung einer HTML-Datei mit einer PDF-Datei bildet einen Graph wie den nachstehenden:

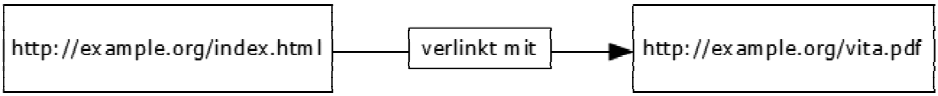


Abb. 15: Verbindung von zwei Dateien im Web

Die einzige Bedeutung, die diesem Graph beigemessen werden kann, ist ein unbestimmter Zusammenhang: die Entität *<http://example.org/index.html>* ist "irgendwie" mit der Entität *<http://example.org/vita.pdf>* verbunden, weil beide etwas (Unbestimmtes) miteinander zu tun haben. Die Bestimmung der Relation beider Entitäten kann im WWW nur implizit und virtuell erfolgen, indem beide Entitäten und der Charakter des sie verbindenden Hyperlinks lesend rezipiert werden und intellektuell miteinander in eine vage Beziehung gesetzt werden:

```
<http://example.org/vita.pdf>
<enthält den Lebenslauf des Autors von>
<http://example.org/index.html>
```

Eine zukünftige Weiterentwicklung des Web im Sinne einer automatisierten Interpretation solcher Relationen wird gegenwärtig heftig diskutiert und eine Reihe verschiedener Technologien und Paradigmen spielen dabei eine gewichtige Rolle. Dabei wird meinungsübergreifend angenommen, dass diese nächste Stufe vor allem durch die effektive Strukturierung von Daten ermöglicht werden wird.

3.2 Das Semantic Web

Die Vision lieferte Tim Berners-Lee 2001 in einem vielzitierten Beitrag im amerikanischen Fachmagazin *Scientific American*. Darin beschrieb er eine zukünftige Informationswelt, in der intelligente Agentensysteme auf der Grundlage strukturierter Daten Entscheidungen treffen und nutzerspezifische Routinearbeiten selbstständig organisieren können. Berners-Lee schrieb:

> *„The Semantic Web […] is an extension of the current [Web], in which information is given well-defined meaning, better enabling computers and people to work in cooperation."* [49]

Grundsätzlich geht es also darum, Daten so zu strukturieren und zu verknüpfen, dass sie mit expliziter Bedeutung versehen sind und durch eine Maschine, einen Computer, einen Algorithmus sinnvoll weiterverarbeitet werden können. Als eine Art *„Idee der zukünftigen Entwicklungsrichtung des WWW"* zielt es darauf ab, *„Webinhalte (zum Beispiel Daten) offener zu gestalten, damit sie nicht nur verarbeitet, sondern auch interpretiert werden können."* [50] Offen bedeutet dabei, dass die strukturierten Daten dabei granular und in standardisierter Form frei im Web verfügbar sind und über einen URI eindeutig identifiziert sind. Da im WWW alle Inhalte über eine URL identifiziert sind, stellen URIs das Bindeglied zum Semantic Web dar.

URI

Die Definition des W3C lautet „The Web is an information space [and] URIs are the points in that space" [51]*– Ein Uniform Resource Identifier ermöglicht die eindeutige Benennung eines Informationsobjekts für das gesamte Internet und bildet den Überbegriff für URLs (ortsabhängige Bezeichner) und URNs (ortsunabhängige Bezeichner). URIs können völlig verschiedene Formen annehmen:*

- `<http://example.org>`
- `<user@example.org:8080>`
- `<urn:isbn:0397004184>`
- `<info:doi/10.1134/S0003683806040089>`
- `<news:comp.infosystems.www.servers.unix>`

[49] [Berners-Lee et al. 2001]

[50] Vgl. [Lassila 2007], 15

[51] Vgl. [W3C 1993]

Diese Vielfalt wird durch zahllose URI-Schemes erzielt, unter denen z.B. *http*, *ftp*, *info*, *news* und auch *z39.50r*/*z39.50s* zu finden sind. Die Schemata sind bei der *Internet Assigned Numbers Authority* (IANA) registriert und decken unterschiedlichste Anwendungsanforderungen ab.[52] Ist ein URI direkt über HTTP auflösbar (dereferenzierbar, d.h. können Informationen über ihn direkt abgerufen werden), entspricht sie einer syntaktisch identischen URL. Da URIs aber in erster Linie für die Identifizierung und Benennung von Entitäten verwendet werden, müssen sie nicht zwangsläufig auch URLs darstellen.

3.2.1 RDF

Die Interpretierbarkeit wird über die formale Konzeptualisierung beliebiger physischer oder abstrakter Entitäten mitsamt ihren definierenden Eigenschaften realisiert. Diese Modelle werden über standardisierte Repräsentationssprachen beschrieben, die auf einem gemeinsamen Datenmodell aufbauen: dem *Resource Description Framework*, kurz RDF. Das Framework wurde 1999 vom *World Wide Web Consortium* (W3C) vorgestellt und bedient sich eines der grundlegenden Modelle von Sprache: jede faktische Information kann anhand der Satzglieder Subjekt, Prädikat und Objekt in einer tripartiten Struktur beschrieben werden, in der ein Prädikat ein Subjekt und ein Objekt miteinander verbindet:

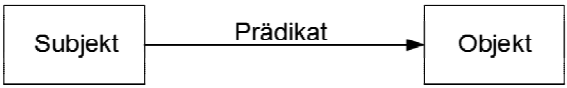


Abb. 16: Aufbau eines Tripels

Ein solches Konstrukt wird als Tripel bezeichnet und bildet das Rückgrat der semantischen Idee. Das Subjekt ist dabei eine Entität, die über eine Prädikatenrelation mit einer anderen Entität verbunden ist. Das Schema eines bibliographischen Objekts hat danach die in [Abb. 17] dargestellte Form.

[52] Vgl. http://www.iana.org/assignments/uri-schemes.html

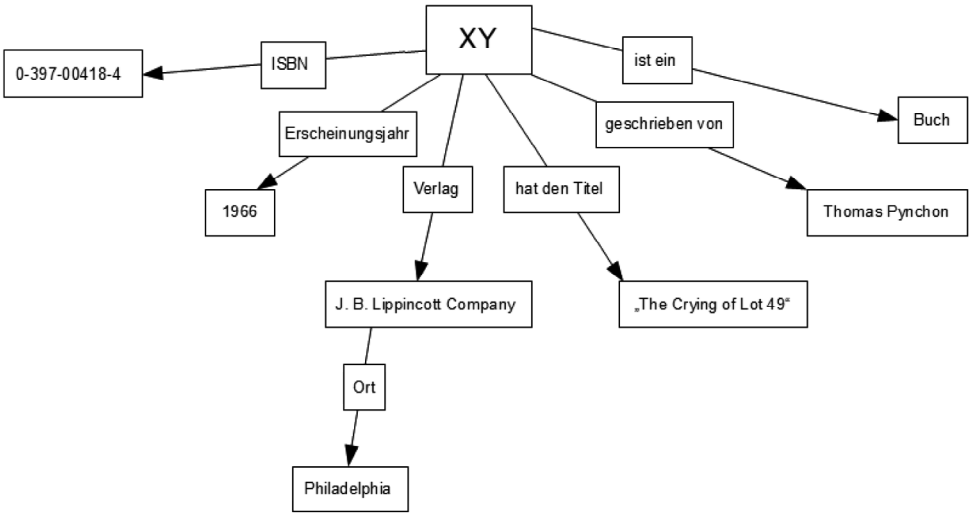


Abb. 17: Schema einer Repräsentation bibliographischer Daten durch Tripelrelationen

Damit eine sinnvolle Verknüpfung erfolgen kann, werden Subjekt und Prädikat immer über einen URI repräsentiert. Das Objekt kann sowohl durch einen URI als auch durch eine Zeichenkette (Literal) repräsentiert werden. Literale werden durch atomare XML-Datentypen normiert und beispielsweise für Datumsangaben verwendet, bei denen ein URI nicht sinnvoll wäre. Die Prädikate werden im semantischen Kontext Properties oder Relationen genannt und entsprechen typisierten Links. Da semantische Modelle „property-oriented" sind und semantische Entitäten nicht in Klassen *„hineingeboren werden, sondern nur durch ihre Properties als Mitglieder einer Klasse verstanden werden"* [53], erhalten sie ihre entscheidende Bedeutung durch die Verknüpfung der Entitäten und ermöglichen – analog zur natürlichen Sprache[54] – erst sinnvolle Aussagen. Ein Beispiel ist das in [Abb. 18] dargestellte Schema: Ein bibliographisches Objekt XY ist durch einen URI benannt und wird durch verschiedene URI-Properties mit einem Wert verbunden, der – wie im Fall des Verlags –

[53] Vgl. [Segaran et al. 2009], 130

[54] Man vergleiche dazu z.B. die Satzkonstruktionen „Ich Pizza" und „Ich mag Pizza".

wiederum ein Subjekt und damit Ausgangspunkt eines Tripels sein kann. Weitverzweigte Graphen – also Aussagenketten – werden so möglich und bilden semantische Netze.

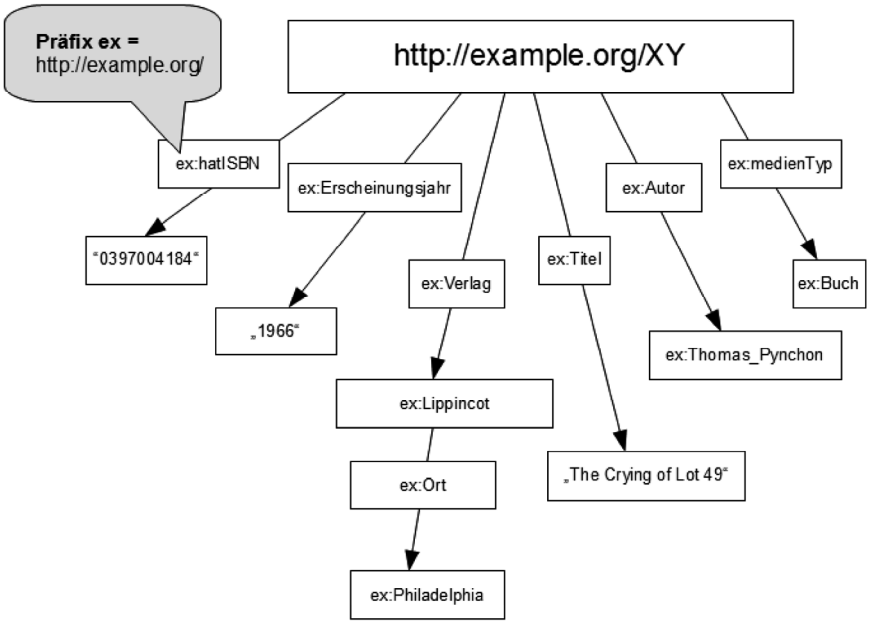


Abb. 18: Schema einer Repräsesentation bibliographischer Daten durch URI-Tripelrelationen

Die Ideen des Semantic Web können in die Ebenen der abstrakten Datenmodellierung und der darauf basierenden Instanzdaten selbst unterteilt werden, die beide auf dem RDF-Paradigma beruhen. Während uns später besonders die Instanzebene beschäftigen wird, soll an dieser Stelle kurz auf die Abstraktionsebene eingegangen werden, die in einem populären Schichtenmodell – dem *Semantic Web Layer Cake* – vom W3C zuammengefasst wurde.[55] Entitäten und Properties werden in Vokabularen definiert, die wiederum auf andere Vokabulare zurückgreifen können. Derartige Vokabulare werden auch als Ontologien bezeichnet und stellen ein mächtiges Werkzeug der Wissens-

[55] http://www.w3.org/2007/03/layerCake.png

repräsentation dar. Da RDF nur das Datenmodell bereitstellt und weder Klassenhierarchien ausdrücken kann noch überhaupt über einen Klassenbegriff verfügt, wurde mit dem *RDF Schema* (RDFS) eine schematische Erweiterung entwickelt.

3.2.2 RDFS

RDFS bildet ein grundlegendes Vokabular und implementiert ein einfaches Klassenmodell zur Modellierung von Entitäten. Es wird über einen eigenen Namensraum *rdfs* integriert (siehe unten) und ermöglicht neben der Einführung von Klassen- und Property-Modellen (*rdfs:Class*, *rdfs:Resource*, *rdfs:Property*) ihre Hierarchisierung (*rdfs:subClassOf*, *rdfs:subPropertyOf*). Properties können zusätzlich auf bestimmte Subjekt- und Objektentitätentypen beschränkt werden (*rdfs:domain*, *rdfs:range*).

Ein einfaches Beispiel: Eine Modellklasse *Buch* und eine Property *hatISBN* werden erstellt und mit der übergeordneten Klasse *Dokument* und der übergeordneten Property *hatIdentifier* in Beziehung gesetzt. Ein *Buch* ist demnach immer auch ein *Dokument* und eine Property *hatISBN* immer auch eine Property *hatIdentifier*. Deren definierte Eigenschaften (z.B. Titel eines Dokuments, syntaktische Vorgaben eines Identifiers) gehen durch diese Relation auf die Unterklassen *Buch* und *hatISBN* über. Dieser Zusammenhang kann über die folgende kodierte Tripelstruktur ausgedrückt werden:

```
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://example.org/Buch> rdf:type rdfs:Class .
<http://example.org/Buch> rdfs:subClassOf <http://example.org/Dokument> .

<http://example.org/hatISBN> rdf:type rdfs:Property .
<http://example.org/hatISBN> rdfs:subPropertyOf <http://example.org/hatIdentifier>

<http://example.org/hatISBN> rdfs:domain <http://example.org/Buch> .

<http://example.org/hatISBN> rdfs:range xsd:integer .
```

Abb. 19: RDFS-Repräsentation

Zunächst werden die benötigten Vokabulare eingebunden. Dazu werden Präfixe eingesetzt, die als Platzhalter für den Namensraum der URI des jeweiligen Referenz-Vokabulars dienen. Die Vokabulare mit den Präfixen *rdf* und *rdfs* enthalten die grundlegenden Vokabulare von RDF und RDFS, in denen die Properties definiert sind, die für die Zuweisung von Objekten zu den Subjekten benötigt werden. Die Entitäten *Buch* und *hatISBN* werden zuerst als Klasse bzw. Property typisiert und von der Klasse *Dokument* bzw. der Property *hatIdentifier* abgeleitet. Abschließend wird die Property *hatISBN* in ihrem Subjekt auf die Klasse *Buch* (weil nur Bücher eine ISBN ausweisen) und das Objekt auf einen ganzzahligen Wert begrenzt (weil ISBN-Nummern bei ausgesparten Trennstrichen aus ganzen Zahlen bestehen). Modelliert man in der Folge ein Objekt vom Typ *<http://example.org/Buch>*, kann auf dieser Grundlage z.B. automatisiert geschlussfolgert werden, dass es sich dabei ebenfalls um ein *<http://example.org/Dokument>* handelt. Außerdem kann abgeleitet werden, dass das Objekt durch seinen Status als *<http://example.org/Buch>* eine ISBN besitzt, deren Wert (im Gegensatz zu anderen Formen von Identifiern) ausschließlich aus ganzen Zahlen besteht.

3.2.3 Serialisierungen

RDF verfügt über mehrere Serialisierungsmöglichkeiten. Während das Turtle-Format und die N3-Syntax vornehmlich auf eine menschliche Rezeption angelegt sind[56], kommt der RDF/XML-Syntax hinsichtlich einer maschinellen Verarbeitung die größte Bedeutung zu. Aus den Tripeln in [Abb. 19] kann das folgende RDF/XML generiert werden:

[56] Die Beispiele dieser Arbeit sind, soweit nicht anders angegeben, in Turtle-Syntax notiert ; Vgl. http://www.w3.org/TeamSubmission/turtle/

```
<?xml version="1.0"?>

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:xsd="http://www.w3.org/2001/XMLSchema#">

  <rdfs:Class rdf:about="http://example.org/Buch">
    <rdfs:subClassOf rdf:resource="http://example.org/Dokument" />
  </rdfs:Class>

  <rdfs:Property rdf:about="http://example.org/hatISBN">
    <rdfs:subPropertyOf rdf:resource="http://example.org/hatIdentifier" />
    <rdfs:domain rdf:resource="http://example.org/Buch" />
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#integer" />
  </rdfs:Property>

</rdf:RDF>
```

Abb. 20: RDF/XML-Repräsentation

Eine wichtige Komponente bilden Namensräume, die die bereits kennengelernten Präfixe abbilden und die eindeutige Bezugnahme auf bestimmte Elemente einer Ontologie nach Wahl ermöglichen. Auf semantischer Ebene erfüllen sie noch eine zweite soziale Funktion, indem sie die Abbildung diverser Gesichts- und Schwerpunkte ein und desselben Konzepts unterstützen. Über die eindeutige URI-Benennung können so die unterschiedlichen Aspekte eines Sachverhaltes über das Web zusammengetragen werden, die dabei durch die Einbindung der verwendeten Vokabulare am Beginn einer jeden RDF-Beschreibung stets nachvollziehbar und zurückverfolgbar bleiben. Durch die URI-Repräsentation werden Subjektentitäten, Properties und Objektentitäten zu offenen, frei im Netz zugänglichen Identitäten deklariert und ist es für eine Maschine möglich, diese in einen Bedeutungskontext mit anderen Entitäten auf Tripelbasis zu stellen.

3.2.4 OWL

Auf Grundlage der bisher behandelten Technologie können Strukturen bereits grob beschrieben werden. Für komplexere Beziehungsformen wie Äquivalenz oder Symmetrie von Entitäten wurde OWL entwickelt. Wie RDFS besitzt OWL einen eigenen Namensraum *owl*. Auch hier ein einfaches Beispiel: Die Klasse *Buch* wird mit der entsprechenden Klasse *Book* eines externen Vokabulars (hier die *Bibliographic Ontology* = *bibo*) gleichsetzend verknüpft. Stößt eine Anwendung auf das lokale Modell *Buch*, kann sie nun über die

owl:sameAs-Relation in der Objektentität weitere Informationen abrufen und automatisch mit der Subjektentität assoziieren.

```
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl:<http://www.w3.org/2002/07/owl#> .
@prefix bibo:<http://purl.org/ontology/bibo/> .

<http://example.org/Buch> owl:sameAs bibo:Book .
```

Abb. 21: OWL-Repräsentation

OWL stellt den derzeitigen State of the Art standardisierter Wissensrepräsentation im Portfolio des W3C dar. Es unterstützt neben ausdrucksstarken Datentypen, Beziehungsformen und Restriktionen auch eine Reihe von Mustern für die automatisierte Schlussfolgerung (Inferenz) und ermöglicht den zukünftigen Einsatz logikbasierter Systeme. Schon in RDFS konnten aufgrund der logischen Fundierung, die einfachen Hierarchiebeziehungen innewohnt, bereits einfache Inferenzen berechnet werden, aber erst in der Ausdrucksstärke von OWL wird diese Fähigkeit soweit ausgebaut, dass sie eine wirkliche Bedeutung erhält. Die *owl:sameAs*-Property aus [Abb. 21] bedeutet beispielsweise in der Sprache der Logik nichts anderes als:

```
IF A == B THEN B == A
```

Weiterführende Informationen zum Thema der logischen Entsprechung von OWL-Elementen und zur Modellierung von Information und Wissen bietet die sehr gelungene Einführung *Semantic Web for the Working Ontologist*.[57]

Stößt OWL an die Grenzen seiner Ausdrucksmittel, kann es durch Regeln erweitert werden. Regeln ermöglichen komplexere Inferenzberechnungen über mehrere Tripel hinweg und können implizites, in den expliziten Daten verborgenes Wissen herleiten. Technisch können Regeln sowohl durch Auszeichnungsformate wie die W3C-Entwicklungen *Rule Markup Language* (RuleML) und *Semantic Web Rules Language* (SWRL) als auch in traditionellen Sprachen wie Prolog formuliert werden. Eine Kompatibilität der verschiedenen Ansätze wird über das *Rule Interchange Format* (RIF) als Austauschformat realisiert. In Version 2 der OWL-Spezifikation sind Regeln fest integriert. Die Fähigkeit zu Schlussfolgerungen bildet die Grundlage für die anspruchsvollen

[57] [Allemang et al. 2008]

Schichten des W3C-Schichtenmodells *Logic*, *Proof* und *Trust*, die die für Verlässlichkeit und Qualitätssicherung von Vokabularen und Daten notwendigen Vertrauensmechanismen bereitstellen sollen, deren ambitionierte Realisierung aber noch in weiter Ferne liegt. Erst mit der Entstehung einer solchen automatisierten Vertrauensinfrastruktur erhält das Semantic Web seine angestrebte Form und Funktion.

3.2.5 Tripelstores und SPARQL

Für die Abfrage von RDF-Daten wurde durch das W3C die Sprache SPARQL entwickelt, die den zielgerichteten Zugriff auf strukturierte Datenentitäten ermöglicht und sich syntaktisch stark an SQL, der Abfragesprache für relationale Datenbanken, orientiert. Im Unterschied zu dieser werden Tripel statt Tupel abgefragt. Damit ein Zugriff möglich wird, werden die Tripel in sogenannten Tripelstores organisiert. Tripelstores sind nichts anderes als auf URIs und kurze Stringliterale optimierte Datenbanken. Die bisher formulierten Tripel aus [Abb. 19] und [Abb. 21] ergeben das Schema in [Abb. 22].

Tripel-ID	Subjekt-ID	Prädikat-ID	Objekt-ID
1	1	1	1
2	1	2	2
3	2	1	3
4	2	3	4
5	2	4	5
6	2	5	6
7	1	6	7

Objekt-ID	Objekt-Wert
1	http://www.w3.org/2000/01/rdf-schema#Class
2	http://example.org/rdf/Dokument
3	http://www.w3.org/2000/01/rdf-schema#Property
4	http://example.org/rdf/Buch
5	http://example.org/rdf/hatIdentifier
6	http://www.w3.org/2001/XMLSchema#integer
7	http://purl.org/ontology/bibo/Book

Subjekt-ID	Subjekt-Wert
1	http://example.org/rdf/Buch
2	http://example.org/rdf/hatISBN

Prädikat-ID	Prädikat-Wert
1	http://www.w3.org/1999/02/22-rdf-syntax-ns#type
2	http://www.w3.org/2000/01/rdf-schema#subClassOf
3	http://www.w3.org/2000/01/rdf-schema#subPropertyOf
4	http://www.w3.org/2000/01/rdf-schema#domain
5	http://www.w3.org/2000/01/rdf-schema#range
6	http://www.w3.org/2002/07/owl#sameAs

Abb. 22: Tabellenschema eines Tripelstores

Möchte man nun abfragen, auf welche externe Entität sich die Äquivalenz der lokalen Klasse Buch bezieht, verwendet man eine Abfrage (*Query*) wie die folgende:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?lokale_resource ?externe_resource
WHERE
{
    ?lokale_resource owl:sameAs ?externe_resource .
}
```

Abb. 23: Abfrage über ein SELECT-Statement in SPARQL

Der Tripelstore wird nach der Tripel-ID abgefragt, in der die Prädikat-ID auf den Wert *owl:sameAs* verweist (Tripel-ID *7*) und liefert das Ergebnis zusammen mit den relationalen Werten für Subjekt-ID und Objekt-ID im SPARQL Query Results Format auf XML-Basis zurück [58] Die Form besteht aus einem <head>-Bereich, in dem die Selektoren der SPARQL-Anfrage aufgeführt sind, und einem <results>-Bereich mit dem resultierenden verknüpften URI *Book* aus dem externen *bibo*-Vokabular.

[58] http://www.w3.org/TR/rdf-sparql-XMLres/

```
<?xml version="1.0" encoding="utf-8"?>

<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <variable name="lokale_resource"/>
    <variable name="externe_resource"/>
  </head>
  <results>
    <result>
      <binding name="lokale_resource">
        <uri>http://example.org/Buch</uri>
      </binding>
      <binding name="externe_resource">
        <uri>http://purl.org/ontology/bibo/Book</uri>
      </binding>
    </result>
  </results>
</sparql>
```

Abb. 24: SPARQL-Antwort

Dies entspricht einem Ergebnis

?lokale_resource	<http://example.org/Buch>
Property	owl:sameAs
?externe_resource	<http://purl.org/ontology/bibo/Book>

SPARQL bietet neben derartigen einfachen Abfragen über den SELECT-Operator auch die Möglichkeit der Konstruktion von RDF-Graphen: Die mächtige CONSTRUCT-Anfrage unterstützt dazu die Definition von Schablonen für die Generierung neuer Tripelinstanzen. Das folgende Beispiel ergänzt die lokale Entität *Buch* um ein Namenslabel aus den verknüpften externen Entitäten:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

CONSTRUCT
{
  ?lokal_resource rdfs:label ?externes_label .
}
WHERE
{
  ?lokale_resource owl:sameAs ?externe_resource .
  ?externe_resource rdfs:label ?externes_label .
}
```

Abb. 25: Generierung eines Tripels über ein CONSTRUCT-Statement in SPARQL

Das neu geschaffene Tripel wird auf einfache Weise in die lokale Informationsbasis integriert:

```
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl:<http://www.w3.org/2002/07/owl#> .
@prefix bibo:<http://purl.org/ontology/bibo/> .

<http://example.org/Buch> owl:sameAs <bibo:Book> ;
<http://example.org/Buch> rdfs:label "Book"@en .
```

Abb. 26: Tripel als Ergebnis des CONSTRUCT-Statements aus [Abb. 25]

Durch solche Prozesse spielt die Abfragesprache eine essentielle Rolle bei der Zusammenstellung neuer Datenkollektionen aus verteilten Datenquellen.

3.2.6 Problemfelder

Zusammenfassend sind die unter dem Begriff Semantic Web subsumierten Technologien der Versuch, das gegenwärtige Web um eine granulare und strukturierte Datenebene zu erweitern, um auf dieser Basis automatisierte Dateninterpretationen zu erreichen. Es hat sich jedoch gezeigt, dass der Begriff bei der Umsetzung einer datenorientierten Netzinfrastruktur eher hinderlich ist und die Diskussion unnötig verbreitert. Die evozierten Missverständnisse über das Wesen und vor allem den Semantikgehalt haben das eigentliche Anliegen der Idee lange verdeckt: die geringe Menge an strukturierten und miteinander verknüpften Daten. Dadurch wurde der Netzwerk-Effekt, der das

WWW so populär gemacht hat, stark ausgebremst und besonders in der ersten Phase der Entwicklung konnte man das klassische „Henne-Ei-Problem" beobachten: durch die geringe vorhandene Datenbasis wurden kaum Investitionen getätigt, was wiederum in einer Zurückhaltung bei der Veröffentlichung strukturierter Daten und bei der Vermittlung der technologischen Zusammenhänge resultierte. Das Semantic Web drohte so zu einem weiteren „Elfenbeinturm" zu werden, zumal auch die Forschungsseite unter der Abwesenheit einer kritischen Masse relevanter Testdaten litt. Nach Jahren des *„scheinbar erfolglosen und akademischen Dahindümpelns"* [59] aufgrund der Datendürre hat in den letzten Jahren ein Umdenken der Forschungscommunity stattgefunden. *„Let's forget about agents zooming around, doing things for you"*, sagte Tim Berners-Lee im Mai 2009 und lenkte den Fokus auf eine praxisorientiertere Lesart:[60] *Linked Data* soll danach das Web mit Daten füllen, indem dort strukturierte und miteinander in Relation gesetzte Information in RDF-Containern veröffentlicht wird.

3.3 Linked Data

„Linked Data lies at the heart of what Semantic Web is all about" [61]

Linked Data geht begrifflich auf ein Memorandum von Berners-Lee aus dem Jahr 2006 zurück, welches er zu einer Zeit veröffentlichte, in der zwar Daten und Ontologien in Ontologien arrangiert wurden, diese aber in ihren originären Domänen (vor allem in Intranets[62]) verhaftet blieben und keinerlei Querverlinkungen zu anderen Datensets aufwiesen. In der Folgezeit forderte er Organisationen und Communities in unzähligen Vorträgen und Workshops mit Nachdruck auf, ihre Daten in großen Mengen zu veröffentlichen und miteinander zu verlinken (*„Raw Data, Now!"*). Er stellte dafür vier einfache Prinzipien vor:[63]

---

[59] [Gradmann 2010]
[60] Vgl. [Berners-Lee 2009], 3
[61] [W3C 2009a]
[62] Vgl. [van Harmelen 2006], 6
[63] Vgl. [Berners-Lee 2006]

1. *Use URIs as names for things*
   Die Entitäten werden durch URIs benannt und identifiziert. Berners-Lee legte fest: *„If it doesn't use the universal URI set of symbols, we don't call it Semantic Web."*
   Beispiel: `<urn:isbn:0397004184>`
2. *Use HTTP URIs so that people can look up those names*
   Die URIs können über die allgemein gültige Webinfrastruktur (HTTP) aufgelöst werden und entsprechen URLs. Damit wird der Gebrauch von URNs ausgeschlossen.
   Beispiel: `<http://example.org/isbn/0397004184>`
3. *When someone looks up a URI, provide useful information, using the standards*
   Bei Auflösung einer URI/URL wird erläuternde Information in standardisierter Form bereitgestellt. Eine Technik namens *HTTP Content Negotiation* versucht zunächst über den HTTP-Header herauszufinden, welche Antwort der Client erwartet und liefert abhängig davon eine maschinen- oder menschenlesbare Form der Information zurück.[64]
4. *Include links to other URIs, so that they can discover more things*
   Die Entitäten werden mit Verknüpfungen zu anderen Entitäten angereichert. Verarbeitende Anwendungen können den Verlinkungen folgen und assoziierte Fakten "entdecken" und integrieren.

Linked Data kann infolgedessen definiert werden als eine Sammlung an *Best Practices* für die strukturierte Verknüpfung von Daten und ihre Veröffentlichung über via HTTP dereferenzierbare URIs zur Schaffung eines globalen und nahtlosen Datenraums auf Grundlage des RDF-Datenmodells:

> *„Linked Data is simply about using the Web to create typed links between data from different sources."* [65]

Die wesentliche Technik dafür ist das Konzept des RDF-Links, das in Erweiterung des Hyperlinks zwischen Dokumenten die Querverlinkung zwischen Daten (sowohl Dokumente als auch RDF-Repräsentationen) auf RDF-Basis (Hyperdaten) ermöglicht. Ein RDF-Link besteht aus einem Tripel, in dem als Subjekt eine URI-Referenz im Namensraum A mit einer URI-Referenz im Namens-

[64] Vgl. Kapitel 4.3.3
[65] [Bizer et al. 2009a], 2

raum B über eine Property (z.B. *owl:sameAs* oder *rdfs:seeAlso*) verbunden ist und bildet die fundamentale Voraussetzung für eine Partizipation der repräsentierten Entitäten im Datenweb. Dezentrale organisierte Daten aus einer Vielzahl von Datenquellen werden über RDF-Links miteinander verwoben.

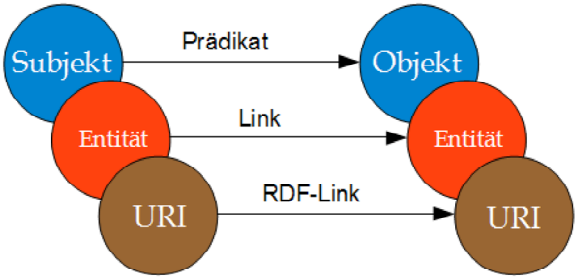


Abb. 27: RDF-Link-Schema

Indem existierende und verbreitete Vokabulare wiederverwendet werden und eigene Modelldefinitionen und Datensammlungen über die Web-Infrastruktur bereitgestellt werden, wird die essentielle Voraussetzung für die Herleitung von Bedeutungszusammenhängen und für die Idee eines Semantic Web geschaffen.

3.4 Linked Open Data

Ein entscheidender Impuls waren die Wechselwirkungen der Linked Data-Community mit der Open Data-Bewegung.[66] Sie führten 2007 zum Linking Open Data-Projekt (LOD) unter der Ägide der *Semantic Web Education and Outreach Group* (SWEO) des W3C. Ziel war es,

> *„to extend the Web with a data commons by publishing various open datasets as RDF on the Web and by setting RDF links between data items from different data sources."* [67]

Bestehende lizenzfreie oder offen lizensierte Daten sollten identifiziert und in das RDF-Datenmodell umgewandelt werden, um sie nach den Prinzipien von Linked Data im Web zu veröffentlichen. Die für die Umwandlung vorgesehe-

---

[66] Zum Beispiel die *Open Knowledge Foundation* (http://www.okfn.org/) und das *Data Portability Project* (http://www.dataportability.org/)

[67] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

nen Daten können dabei in einer beliebigen strukturierten Form vorliegen, ob in relationalen Datenbanken, als kommaseparierte Listen (CSV), in proprietären XML-Formaten oder sonstiger Containergestalt. Die Konversion erfolgt über sogenannte Wrapper, die in der Lage sind, vorhandene Datenstrukturen automatisiert zu extrahieren und auf das RDF-Datenmodell abzubilden. Linked Open Data stellt das sichtbarste Beispiel der Annahme und Anwendung der Prinzipien von Berners-Lee dar und wird mittlerweile oft begrifflich mit Linked Data gleichgesetzt.

Zwei elementare Entwicklungen für die Genese relationaler Datenbankinformation im Datenweb sind der *D2R Server* [68] und das PHP-basierte Framework *Triplify*.[69] D2R Server erlaubt den direkten Zugriff auf Datenbankdaten als Linked Data über eine SPARQL-Schnittstelle. Geeignet vor allem für überschaubarere Informationsmengen, müssen Daten auf diese Weise weder synchronisiert noch doppelt gepflegt werden. Die Entwicklung geht bis 2003 zurück, als mit D2RQ eine XML-basierte Mappingsprache für die Abbildung von Datenbankstrukturen auf das RDF-Datenmodell entwickelt wurde. In den folgenden Jahren wurde der Ansatz zu einer Serverlösung ausgebaut, der in kleineren Produktionsumgebungen eingesetzt werden kann.[70] Das Triplify-Framework ist eine jüngere Entwicklung und zielt auf eine weitergehende Absenkung der Einstiegsbarriere in die semantische Datenrepräsentation ab.[71] Dazu werden unter Verwendung von SQL vordefinierte Mappings auf Datenbankschemata populärer Webanwendungen (Foren, Weblogs, Wikis, Content Management Systeme) bereitgestellt. Das auf der Skriptsprache PHP basierende Framework wandelt in einer Datenbank existierende Werte unter Verwendung existierender Vokabulare in RDF-Tripel um und ermöglicht den Datenzugang über eine SPARQL-Schnittstelle. Ein Härtetest wurde kürzlich in Form der Umwandlung sämtlicher Daten der Graswurzel-GPS-Intiative OpenStreetMap und ihrer Veröffentlichung als *Linked GeoData* erfolgreich bestanden.[72] Weitere Beispiele für die semantische Bereitstellung existierender

---

[68] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

[69] http://www.triplify.org

[70] Für größere Datenmengen eignet sich der Virtuoso Universal Server der Firma OpenLink, vgl. http://www.openlinksw.com/virtuoso/Whitepapers/html/rdf_views/virtuoso_rdf_views_example.html

[71] Vgl. [Auer et al. 2009], 1

[72] Vgl. [Auer et al. 2009], 8f ; Vgl. http://linkedgeodata.org/

Daten sind das XLWrap-Projekt zur Umwandlung von im Excel-Format gespeicherten Tabellen [73] und verschiedene API-Wrapper um populäre Datenlager wie Flickr, Amazon und diverse soziale Netzwerke.[74] Die durch solche Prozesse und Methoden aggregierte Menge an semantisch ausgezeichneten Daten und ihre Konstellation zueinander bildet die *LOD Cloud* im Web.[75] Anfangs vor allem durch Akademiker und Entwickler aus dem universitären Bereich sowie kleineren Firmen vorangetrieben, waren schon 2008 größere Organisationen wie die Library of Congress, Thomson Reuter und die BBC mit einer Reihe von Projekten signifikant beteiligt und der offene Charakter von LOD hat in der Folge für anhaltendes Wachstum und weitere namhafte Beteiligungen aus unterschiedlichsten Domänen wie Wissenschaft, Forschung, Kultur und eGovernment-Programmen gesorgt, die dazu beigetragen haben, das Datenweb mit hochrelevanten und granularen Informationseinheiten zu bevölkern.[76] In radikaler Erweiterung des Mashup-Prinzips des „Web 2.0" und der Überwindung der beschränkten Zugriffsmöglichkeit auf spezifische API-Schnittstellen stehen dabei sämtliche Datenquellen gleichzeitig und einheitlich zur Abfrage über offen standardisierte Schnittstellen zur Verfügung: die Schaffung einer ultimativen API.

---

[73] http://xlwrap.sourceforge.net/

[74] Eine ausführliche Aufstellung von Konvertern ist unter http://esw.w3.org/topic/ConverterToRdf abrufbar.

[75] Vgl. [Abb. 28] ; aktuelle Versionen sind unter http://linkeddata.org verfügbar.

[76] Eine stetig aktualisierte Liste beteiligter Institutionen ist abrufbar unter http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/NewsArchive

4 Bibliotheken und Linked Data

4.1 Warum Linked Data?

Emmanuelle Bermès von der französischen Nationalbibliothek beschreibt ein bibliothekarisches Engagement im Datenweb als „Gelegenheit, ein Netzwerk des Wissens aufzubauen, das auf akzeptierten Standards beruht und die Bedürfnisse einer breitgestreuten Nutzerschaft bedient".[105] Damit Bibliotheken dort eine signifikante Rolle spielen können, ist eine Kehrtwende in der Distribution ihrer Metadaten notwendig. Diese wurden über Jahrzehnte hinweg in geschlossenen Datenbanken – „Datensilos" – vorgehalten und für die Öffentlichkeit über recherchierbare Kataloge zugänglich gemacht. Dort können über definierte Benutzerschnittstellen Suchanfragen an den Server gestellt werden (in der Regel über HTTP-POST) und dieser antwortet mit einer dynamisch generierten Ergebnisliste. Die dynamisch generierten Seiten werden über populäre Suchmechanismen wie Google aber nicht indexiert und so sind die Daten für das WWW praktisch unsichtbar. Einzig die Kataloge selbst sind auffindbar, die Daten dagegen Teil des sogenannten *Deep Web*. Die indirekte Bereitstellung über Kataloge reicht also nicht aus und wird dem Wert der Daten nicht gerecht. Martin Malmsten, IT-Entwickler in Diensten des schwedischen Verbundkatalogs LIBRIS, hat das Problem 2008 in einem Beitrag auf der Mailingliste NGC4LIB pointiert erfasst:

> *"For too long have library data been trapped within data-silos only accessible through obscure protocols. Why is access to library data still an issue? This was solved in a matter of months on the web, when Excite (or whichever search engine was first) was introduced. Why are there not at least ten search engines containing the majority of the world's bibliographic data?"* [106]

Die Schaffung eines direkten Zugangs durch die offenen und transparenten W3C-Standards des WWW ist daher die grundlegende Voraussetzung für die Auffindbarkeit.[107] Auch eine geschäftsmäßige Verwendung der Daten durch externe Informationsanbieter, Entwickler, Wissenschaftler und andere Fach-

---

[105] Vgl. [Bermès 2009], 1

[106] [Malmsten 2008b]

[107] Organisationen wie OCLC ermöglichen zwar die Einbindung in Suchergebnisse von Google, jedoch nur durch explizite Vereinbarungen mit dem Suchmaschinenbetreiber.

gemeinden ist durch den technologischen Status Quo stark eingeschränkt. So ist es beispielsweise für externe Entwickler fast unmöglich, eine ausreichende Menge miteinander in Relation stehender bibliographischer Daten zusammenzutragen, um einen Anwendungsprototypen, der diese Daten weiterverarbeiten soll, auf sein *Proof of Concept* hin zu überprüfen.

Rupert Hacker schreibt noch 2000 in seinem Bibliothekarischen Grundwissen, dass bei Verfahren zum Austausch von bibliographischen Daten stets vorausgesetzt werden muss,

> *"dass sendende und empfangende Bibliotheken das gleiche Kategorienschema oder Datenformat für die bibliographischen Elemente der Titelaufnahme verwenden."* [108]

Unter Voraussetzungen wie diesen erscheint eine interdisziplinäre Kollaboration geradezu utopisch, da der für den Datenaustausch dominierende Standard MARC einschließlich seiner Derivate und Ausformungen außerhalb der bibliothekarischen Domäne nur wenig Resonanz findet. Zwar behauptet [Chowdury 2007], dass das Format von einer breiten Informationscommunity *„begeistert angenommen"* wurde[109], die Befunde des Stillwater-Reports von 2004 und des 2008 veröffentlichten Berichts *On the Record* kommen indes zu einer völlig anderen Einschätzung der Situation: die Handhabung bibliographischer Daten mit MARC sei *„out of step"* mit aktuellen Technologien [110] und *„incompatible with current database principles"* [111] und werde aufgrund dieser Inkompatibilitäten von externen Fachgemeinschaften und deren bibliographischen Anwendungen gemieden. Vielmehr stelle es – wie [Tennant 2002] treffend formulierte – einen *„geheimnisvoll-obskuren Standard"* dar, der nur über geringe Reichweite außerhalb der Bibliothekswelt verfügt.[112] Das Format sei daher

> *"not an appropriate starting place for a new bibliographic data carrier because of the limitations placed upon it by the formats of the past."* [113]

Dass Linked Data dabei eine ernsthafte Alternative darstellt, hat [Bermès 2009] veranschaulicht:

---

[108] [Hacker 2000], 218
[109] Vgl. [Chowdury 2007], 53
[110] Vgl. [LC Working Group 2008], 24
[111] [Stillwater 2004], 7
[112] Vgl. [Tennant 2002]
[113] Vgl. [LC Working Group 2008], 24

> *„It's just as if you could link a MARC datafield inside a record from your catalogue with a Dublin Core element inside a record in another catalogue, and then use this link to retrieve and manipulate both records together without the need for any metadata mapping / crosswalk, record exchange / import, or software development..."* [114]

Die Möglichkeit, bibliographische Daten in einem gemeinsamen Datenmodell miteinander in beliebige Beziehungen setzen zu können, bietet die notwendigen Voraussetzungen, die bisherigen Barrieren der externen Datenkommunikation abzutragen. Auch Suchmaschinen können auf diese Weise bibliographische Daten indexieren und in ihren Suchergebnissen verarbeiten. Das URI-Prinzip von Linked Data sorgt dafür, dass die beschriebenen bibliographischen Entitäten eindeutig verlinkbar sind und nicht in den Katalogen verborgen bleiben. Eine Adaptierung bietet eine Reihe nützlicher Nebeneffekte:

- Bibliographische Daten werden durch wesentlich flexiblere Retrievalsprachen in der Tradition von SQL nach völlig neuen Kriterien durchsuchbar. Abfragen à la "Finde alle Großstadtautoren aus Asien" oder gar "Finde alle literarischen Erstlingswerke von Autoren, die nicht älter als 40 Jahre alt wurden", werden bei entsprechender Datenbasis trivial.
- Bibliotheken sind nicht mehr "allein" bei der Bewältigung zukünftiger Herausforderungen, sondern können im Dialog mit der Webgemeinde an übergreifenden und zukunftsfähigen Lösungen arbeiten.
- Bibliotheken erhalten neue und nicht-bibliothekarische Perspektiven auf ihre Daten, die der "realen Welt" entsprechen. Eine direkte und offene Veröffentlichung der Daten lässt Bibliotheken von entstehenden Datenmixturen und Anwendungen profitieren, die weit über die Ergebnisse sogenannter Hackathons und Mashatons, wie sie z.B. von OCLC veranstaltet werden, hinausgehen.[115]Auf wirtschaftlicher Ebene – nicht zuletzt in Zeiten knapper Etats im Bibliothekssektor [116] – können die finanziellen Mittel, die durch den Abbau notwendiger Datenkonvertierungen und die Abkehr von kostenintensiver Spezialsoftware und Schnittstellen für den Import und Export externer Datenbestände in ein laufendes Bibliothekssys-

[114] [Bermès 2009], 2

[115] Vgl. http://worldcat.org/devnet/wiki/Presentations ; Norbert Weinberger von OCLC Deutschland bemerkte am 28.01.2010 in einem Vortrag am *Institut für Bibliotheks- und Informationswissenschaft* der Humboldt Universität zu Berlin, dass als Ergebnis dieser Veranstaltungen nicht selten neue OPACs entstanden.

[116] Vgl. u.a. [Deutscher Kulturrat 2009]

tem eingespart werden[117], an anderer Stelle eingesetzt werden, etwa bei der (digitalen) Verfügbarmachung einzigartiger und besonderer Artefakte [118] oder bei der Entwicklung zeitgemäßer Katalogisierungsregeln.

- Bibliotheken können endlich technologische Unabhängigkeit gegenüber den Anbietern von Bibliothekssoftware erlangen. Die Auswahlmöglichkeiten von Bibliotheken waren jahrzehntelang aufgrund hochspezialisierter Anforderungen auf Systeme beschränkt, die MARC und bibliotheksspezifische Schnittstellen wie z39.50 unterstützen. Offene Webtechnologien ließen Bibliotheken in erheblichem Maße von gemeinschaftlichen Entwicklungen der Open Source-Bewegung profitieren und zugleich dem paralysierenden „Würgegriff" bisheriger Systeme und ihrer Anbieter entkommen, welcher im Laufe der Jahre für ein großes Defizit an technischer Flexibilität gesorgt und damit die zeitnahe Hinwendung zu technischen Alternativen verhindert hat.[119]

Linked Data verspricht auch auf theoretischer Ebene wichtige Impulse. Die Abwesenheit eines allgemeinen Konzeptmodells für bibliographische Entitäten hat zu gewaltigen Inkonsistenzen im Datenbestand geführt. Bereits seit 1998 steht mit den *Functional Requirements for Bibligraphic Records* (FRBR) ein konzeptuelles Modell zur Verfügung, das Klassen und Relationen für ein explizites Schema bibliographischer Entitäten einführt und deren eindeutige Lokalisierung im bibliographischen Kosmos verspricht.

FRBR

FRBR wurde 1998 als Entity-Relationship-Modell konzipiert und von der IFLA mit dem Ziel entwickelt, eine interoperable theoretische Grundlage für bibliographische Daten zu schaffen. Das Modell stellt eine „semantic expression of the relationships between items in the library catalog" dar.[120] Neben textuellen Objekten sind auch audiovisuelle Entitäten interpretierbar. FRBR ermöglicht die Typisierung und eindeutige Lokalisierung jedes bibliographischen Objekts im bibliographischen Kosmos und besitzt durch seine Attribute/Properties eindeutige Relationen zu anderen Objekten. Es bildet in drei Gruppen einen konzep-

[117] Beispielsweise http://www.exlibrisgroup.com/de/files/Germany/LizenzpflichtigeALEPHAddons.pdf

[118] Vgl. [LC Working Group 2008], 21ff

[119] Vgl. [Gradmann 2005], 64

[120] [Wallis 2004]

tuellen Rahmen, in dem bibliographische Entitäten und ihre Beziehungen zueinander sowie Relationen mit ihren verantwortlichen Produzenten und ihren thematischen Bezügen definiert sind:

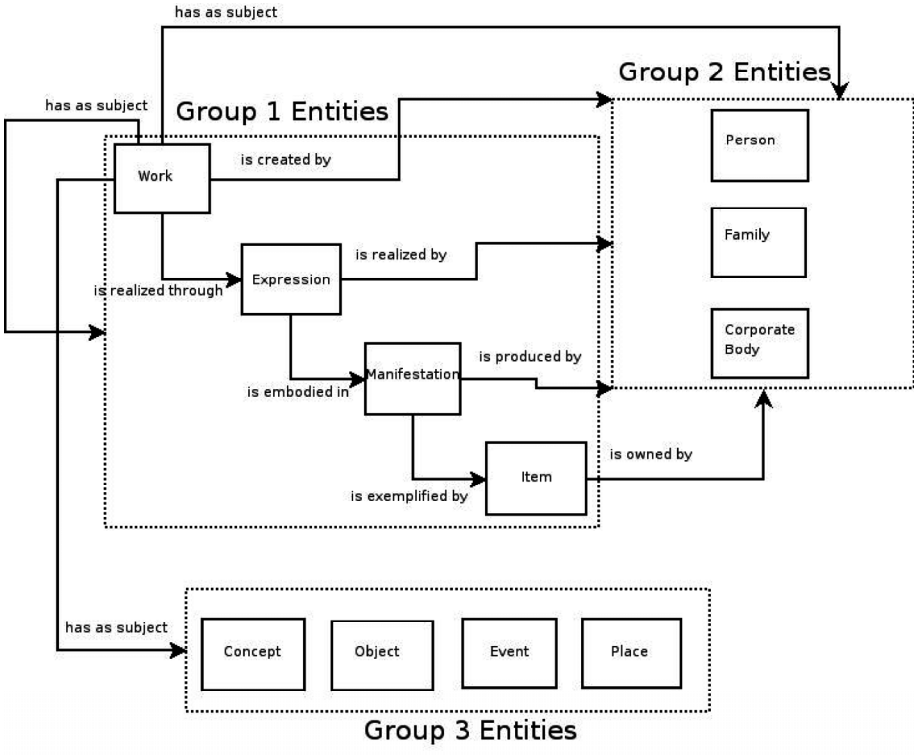


Abb. 33: Beziehungen von FRBR-Entitäten [121]

FRBR wird flankiert von den Functional Requirements for Authority Data (FRAD) und den Functional Requirements for Subject Authority Records (FRSAR). Beide sind bislang kaum getestet.

Die Bedeutung eines konzeptuellen Modells für die Analyse von Systemen und ihrer expliziten Prozesse und Relationen ist allgemein anerkannt und hat gravierende (positive) Auswirkungen auf ihre Effizienz. Schon Computerpionier Douglas Engelbart schrieb 1962:

[121] Zitiert nach: http://www.frbr.org/files/entity-relationships.png

> „The conceptual framework […] must orient us toward the real possibilities and problems associated with using modern technology to give direct aid to an individual in comprehending complex situations, isolating the significant factors, and solving problems." [122]

Über ein Jahrzehnt nach seiner Entwicklung befindet sich FRBR noch immer in einem Stadium der Evaluation, da eine Implementierung durch die bestehende Dateninfrastruktur erschwert wird. Das Modell ist nicht unumstritten in seiner Eignung für nicht-physische, d.h. digitale Objekte und lässt Fragen offen, die vor allem abstrakterer Natur sind: Was konstituiert ein Werk? Ist ein Zeitschriftentitel ein Werk? [123] *Wie können atomarere Entitäten wie Kapitel oder Zeitschriftenartikel typisiert und eingeordnet werden? Über die Verknüpfungspotentiale von Linked Data und eine große Menge an triplifizierten bibliographischen Daten könnten die Konzepte von FRBR endlich ausgiebig und vor allem effektiv getestet werden.*

Die Frage lautet also:

Wie können Bibliotheken Linked Data produzieren?

Der Bericht *On the Record*, der 2008 von der *Working Group on the Future of Bibliographic Control* veröffentlicht wurde, hat die Agenda für eine Überführung bibliographischer Daten in eine interoperable Form bereits umrissen:

> *„Express library standards in machine-readable and machine-actionable formats, particular those developed for use on the Web […] Provide access to standards through registries or Web sites so that the standards can be used by any and all Web applications"* [124]

Die kontrollierten Vokabulare, die bislang zum Teil noch in Papierform oder in proprietärer elektronischer Form vorliegen[125], sollen in eine maschinell-lesbare und interoperable Form umgewandelt werden. Anschließend können bibliographische Titeldaten mit diesen semantischen Normdaten-Entitäten angereichert werden. In einer Stellungnahme der Open Knowledge Foundation zu den Ergebnissen der Studie heisst es dazu:

---

[122] [Engelbart 1962], II

[123] Vgl. [Delsey 2003]

[124] [LC Working Group 2008], 25

[125] Z.B. über die DNB-BEnutzerschnittstelle *Melvil* (http://melvil.d-nb.de/swd) oder das *Classification Web* der Library of Congress (http://classificationweb.net)

"Bibliographic records are a key part of our shared cultural heritage. They too should therefore be made available to the public for access and re-use without restriction. Not only will this allow libraries to share records more efficiently and improve quality more rapidly through better, easier feedback, but will also make possible more advanced online sites for book-lovers, easier analysis by social scientists, interesting visualizations and summary statistics by journalists and others, as well as many other possibilities we cannot predict in advance." [126]

4.2 Theoretische Vorüberlegungen

Wie könnten bibliographische Daten in Form von Linked Open Data aussehen? Mit der Wahl aussagekräftiger Bezeichner für die zu beschreibenden Entitäten und einer Analyse der zur Verfügung stehenden Vokabulare zur Annotation sind dafür zunächst zwei Fragen hinsichtlich einer effektiven zukünftigen Nutzung der Daten zu erörtern.

4.2.1 Die Identifizierung der Entitäten durch HTTP-URIs

Im Zentrum steht die Frage nach sinnvollen und stabilen Identifiern für die beschriebenen Entitäten. Nach den ersten beiden Prinzipien von Linked Data kommt dafür nur die Verwendung des HTTP-URI-Schema in Frage.[127]

4.2.1.1 Sinnvolle Identifizierung

Ein identifizierender URI sollte auf implementierende Details wie spezifische Ports, Serververzeichnisse, Parameter und sonstige physische Pfadangaben verzichten und stattdessen über serverseitige Mappings die Struktur des jeweiligen Datenraums angemessen abbilden. Von einem URI der Variante

```
<http://example.org:8080/rdfizer/cgi-
        bin/rdf.php?id=Pynchon>
```

ist aufgrund einer schwierigen Referenzierbarkeit und von Sicherheitsaspekten sowie vor allem aus Gründen technologischer Flexibilität unbedingt abzuraten. Stattdessen sollte ein URI-Muster á la

```
<http://example.org/rdf/Pynchon>
```

[126] [Open Knowledge Foundation 2008]
[127] Vgl [Bizer et al. 2007], 3

gebildet werden, welches serverintern auf jeden beliebigen Pfad abgebildet werden kann (z.B. über das Modul mod_rewrite für den Apache Webserver), nach außen aber kurz und prägnant erscheint. Auf diese Weise können eingesetzte Technologien durch geringfügige Anpassungen in der Serverkonfiguration ausgetauscht werden, ohne bereits referenzierende Vokabulare und Datensets „ins Leere laufen zu lassen". Ein sinnvolles Muster ist

```
<http://[DOMAIN]/resource/[Entitätentyp]/[Identifier]>
```

Der Entitätentyp bildet einen Namensraum und hat dabei entweder die virtuelle Pfadangabe */bib/* für die Benennung einer bibligraphischen Einheit oder */auth/* für normierte Entitäten wie Personen und Schlagworte. Um eine besitzende Institution anzuzeigen, kann */library/* angegeben etc. Eine Eindeutigkeit meint, dass

> *„no matter where in the universe a URI is used, a specific URI represents one and only one resource"* [128]

Der Rückgriff auf spezifische Identifier aus einem zugrundeliegenden System – z.B. die mit den Datensätzen korrespondierenden Kontrollnummern im ILS – unterstützt diesen Anspruch. Eine alternative Verwendung von automatisch zugewiesenen textuellen Bezeichnern – z.B. nach dem Muster Autor_Titel – kann die menschliche Rezipierbarkeit unterstützen, birgt aber stets die Gefahr von mehrdeutigen URIs, die durch metrische und statistische Softwaremechanismen aufgelöst werden müssen.

4.2.1.2 Stabile Identifizierung

Die Stabilität eines URI stellt sicher, dass eine Entität konsistent und dauerhaft verwendet werden kann:

> *"A URI today should represent the same resource tomorrow."* [129]

Damit wird das Persistenzproblem adressiert, das in der Vergangenheit vor allem im Bereich digitaler Repositories und deren Bestrebungen, digitale Datenobjekte technologisch unabhängig und dauerhaft identifizieren zu können, diskutiert worden ist. Unterschiedliche dedizierte Ansätze wurden verfolgt, um die Identifizierung solcher Objekte einer stärkeren (institutionellen) Kontrolle zu unterwerfen, die vor allem auf dem Einsatz bestimmter URI-

[128] [Segaran et al. 2009], 109
[129] Ebd.

Schemata wie *Document Object Identifiers* (DOIs) oder URNs fußen. Diese Ansätze bergen drei Schwächen:

a) Die Namensräume der Schemata müssen notwendigerweise durch durch zentrale und nicht-verteilte Resolver, deren Verfügbarkeit stets gewährleistet sein muss, dereferenziert werden. Fallen diese aus, sind die Identifier nicht mehr in die entsprechende URL auflösbar und die gewünschte Persistenz hinfällig.
b) Die Informationsinfrastruktur wird zum Teil unnötig zersplittert.
c) Es existiert eine große Zahl allgemein verwendeter, jedoch nicht standardisierter URI-Schemata.[130]

In Bezug auf eine technologische Durchsetzung von persistenten Identifikationsmechanismen hat Tim Berners-Lee 1998 in seinem Memo *Cool URIs don't change* darauf hingewiesen, dass die Persistenz von URI-Identifiern in erster Linie ein durch Menschen verursachtes und kein technisches Problem sei:

> *"URIs don't change: people change them."* [131]

Persistenz per se sei demnach ein Konstrukt und die Auswahl dauerhaft gültiger URIs demnach vielmehr eine „soziale Herausforderung“[132] und könne nur durch einen Konsens innerhalb einer Community erreicht werden. Die Auswahl und das „Design“ eines kurzen, sinnvollen und quasi-persistenten URI-Musters ist damit sowohl eine technische als auch eine stilistische Herausforderung.[133]

4.2.2 Die (Wieder-)Verwendung bibliotheksrelevanter Vokabulare

Für die Beschreibung bibliographischer Daten können sowohl die bereits vorgestellten als auch bibliotheksspezifische Vokabulare verwendet werden. Die Flexibilität des RDF-Datenmodells erlaubt die gleichzeitige Verwendung verschiedener Vokabulare. Das hat den Vorteil, dass bei Abwesenheit geeigneter Properties oder Klassen auf einfache Weise Definitionen eingebunden und nahtlos verwendet werden können.

---

[130] Vgl. dazu http://en.wikipedia.org/wiki/URI_scheme#Unofficial_but_common_URI_schemes
[131] [Berners-Lee 1998]
[132] Vgl. [Sasaki 2008], 28-29
[133] Vgl. [Malmsten 2009], 2

Für die Annotation von personaler Information von Autoren und anderen Beitragenden eignet sich das bereits beschriebene FOAF-Vokabular bestens. Handelt es dabei um Personen der Vergangenheit, können deren Lebensdaten über die Bio-Ontologie erfasst werden. Ein Autor kann wie folgt definiert werden:

```
@prefix base:<http://example.org/> .
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf:<http://xmlns.com/foaf/0.1/> .
@prefix owl:<http://www.w3.org/2002/07/owl#> .
@prefix bio:<http://purl.org/vocab/bio/0.1/> .

<AutorXY>
  rdf:type foaf:Person ;
  foaf:name "Thomas Pynchon" ;
  foaf:givenName "Thomas Ruggles Jr." ;
  foaf:surname "Pynchon" ;
  foaf:gender "Male" ;
  bio:birth "1937-08-05" ;
  owl:sameAs <http://dbpedia.org/resource/Thomas_Pynchon> ;
  :profession <http://dbpedia.org/resource/Writer> .
```

Abb. 34: Annotation via FOAF und Bio

Um die einfache und flexible Erweiterbarkeit zu demonstrieren, wird zusätzlich über *owl:sameAs* eine Äquivalenz mit der entsprechenden DBpedia-Entität des Autors hergestellt und über eine eigene Property *foo:profession* die Tätigkeit des Autors auf die externe Klasse *Writer* abgebildet.

Die im weiteren Verlauf zu besprechenden bibliotheksspezifischen Vokabulare können in drei Kategorien eingeteilt werden: die Beschreibung bibliographischer Metadaten (DublinCore, Bibo), normierter Daten (SKOS) und bibliographischer Beziehungen gemäß FRBR (FRBRcore, RDA).

5 Ausblicke und Fazit

Bibliotheken werden über Linked Data endlich und wahrhaftig mit dem Web verbunden.[186] Die semantische Beschreibung von bibliographischen Daten durch das einheitliche RDF-Datenmodell schafft eine Alternative zur historisch gewachsenen Zersplitterung der bibliothekarischen Dateninfrastruktur. Ohne Rücksicht auf Schemakompatibilitäten können sie miteinander integriert werden. Nachdem die Bibliothekswelt es in der ersten Hochphase des Internets versäumt hat, sich als Anbieter hochqualititaver Webinformation zu positionieren, erhält sie eine weitere Gelegenheit, einen wichtigen Part der zukünftigen Informationsinfrastruktur besetzen zu können. Bibliotheken bietet sich die Möglichkeit, das Fundament bibliographischer Daten auf eine neue Ebene zu heben und den Dialog mit der Webgemeinde zu intensivieren. Linked Data sollte dabei nicht als *eine zusätzliche Technologie* wahrgenommen werden, sondern stattdessen als eine zweite Chance der Bibliotheken im Web begriffen werden. Bibliotheken haben sich durch ihr akribisches und hoch qualitatives Schaffen über Jahrzehnte und Jahrhunderte einen Status der vertrauenswürdigen und verlässlichen Vermittlung von Wissen erworben. Angesichts einer angestrebten Erneuerung dieser exponierten Stellung im Informationszeitalter und im Kontext eines allgegenwärtig verfügbaren Web, das stetig mehr Information unterschiedlichster Qualität anbietet, müssen sie sich daher als relevante Instanzen des Wissens positionieren. Eine wesentliche Kernaufgabe wird dabei das „relationship business" sein, die Konglomeration von zusammenhängender und gesicherter Information jeder Art und ihre Überführung in Wissen.[187] Das beginnende 21. Jahrhundert wird von [Marcum 2008] als "crucial time for reshaping bibliographic activity" angesehen, in der die Anwendung neuer Theorien und ihrer Evaluation Priorität habe.[188] Verlinkte und offene bibliographische Bibliotheksdaten bieten dafür nach [Bermès 2009] ein ideales Einsatzszenario, weil "kurzfristige Ergebnisse möglich" seien.[189]

---

[186] Vgl. [Malmsten 2009], 1
[187] Vgl. [Eversberg 2010]
[188] Vgl. [Marcum 2008], 4
[189] [Bermes 2009], 1

5.1 Strategien

Zwei unterschiedliche Strategien im Umgang mit Linked Data werden gegenwärtig verfolgt: *Structure First* und *Data First*. Der erste Ansatz geht davon aus, dass Strukturen geschaffen werden müssen, um Daten adäquat aufzufangen. Die bisherige Wirkungsgeschichte der Semantic Web-Idee spricht allerdings gegen einen solchen Ansatz. Fehlende Daten waren es, die eine Praxistauglichkeit der Idee lange Zeit verhinderte und erst erfolgreiche Initiativen wie Linked Open Data führten zu einer Belebung. Der Ansatz von *Data First* dagegen entspricht dem iterativen Ansatz in den LIBRIS- und LCSH-Projekten. [Mazzocchi 2005] hat dieser Strategie eine "effizientere Nutzbarkeit" zugebilligt und drei Gründe dafür aufgeführt:[190]

- *Data First* entspricht dem natürlichen Lernen und der Entwicklung von Sprache. Regeln, Modelle, Abstraktionen und Kategorien werden erst gebildet, nachdem Information und Daten zur Verfügung stehen:

  "This is why it's easier to learn a computer language from examples than from its theory, or a natural language by just being exposed to it instead of knowing all rules and exceptions."

- Die Komplexität des Systems wird langsam aufgebaut und kann leichter optimiert und an die bestehenden Umstände angepasst werden.
- Erträge sind unmittelbar wahrnehmbar und können einfacher ausgebaut werden.

Beide strategischen Richtungen manifestieren sich in der Unterscheidung von Linked Data und Linked Open Data. In den Aktivitäten der großen Bibliotheksorganisationen wird Linked Data als neues Datenparadigma anerkannt, resultiert aber nur selten in der Veröffentlichung konkreter Instanzdaten. Die Entwicklung von RDA ist die klassische Variante des strukturellen Ansatzes, indem ein komplexes Modell aus Entitäten und Relationen aufgebaut wird, in das später Daten eingepflegt werden sollen. Dieses Paradigma herrscht vor allem in den Bereichen vor, in denen die Kontrolle über Metadaten einen wirtschaftlichen Nutzen erfüllt. Kommerziell ausgerichtete Einrichtungen wie OCLC mit restriktivem Geschäftsmodell nutzen Linked Data für die interne Organisation ihrer Daten, gedenken aber auch zukünftig, ihre Einkünfte aus dem Verkauf bibliographischer Information juristisch abzusichern.[191] [Yang et

[190] Vgl. [Mazzocchi 2005]
[191] Vgl. u.a. [Pohl 2009]

al. 2009] verweist auf eine Reihe weiterer infrastruktureller Projekte aus den Bereichen Digitale Bibliotheken und Repository-Software, in denen die Strukturen für einen Einsatz von Linked Data vorbereitend realisiert werden.[192]Infrastrukturell unabhängig, zielt die Strategie von *Data First* auf eine freie Veröffentlichung von Metadaten als (Linked) Open Data ab, um aus ihnen die passenden Strukturen abzuleiten. Dafür ist die freie und massenhafte Veröffentlichung gemeinfreier Metadaten Grundvoraussetzung. Die Form ist dabei zunächst zweitrangig. Die Veröffentlichung ist in erster Linie ein "politisch-rechtlicher Prozess"[193], in der Einsicht, dass die durch Bibliotheken hergestellten Metadaten Erzeugnisse einer öffentlichen Steuerfinanzierung darstellen und daher der allgemeinen Öffentlichkeit wiederum uneingeschränkt und offen zur Verfügung stehen sollten.

Jüngst haben einige bibliothekarische Einrichtungen beispielhaft den Weg von Open Data beschritten: nachdem bereits im Dezember 2008 die Bibliothek der Universität Huddersfield (USA) 80.000 Datensätze freigab [194], veröffentlichte ein Jahr später die Bibliothek des Genfer Forschungszentrums CERN ihre Metadaten als MARCXML und im März 2010 wurden sämtliche Katalogdaten der Universitäts- und Stadtbibliothek Köln in Kooperation mit dem hbz zum Download bereitgestellt.[195] Die Freigaben stießen innerhalb und außerhalb der Bibliotheksdomäne auf große Resonanz und erste kleinere Nachnutzungen folgten kurze Zeit später.[196] Der Triplifizierung und Wiederbereitstellung der Daten werden die kommenden Monate gelten. Die dadurch gewonnenen semantischen Repräsentationen bieten beste Testobjekte für bibliographische Anwendungen und Webservices.

Die Strategien der Nationalbibliotheken sind im Vorfeld der RDA-Einführung und angesichts eines sich wandelnden Marktes für bibliographische Information kaum zu deuten. Die DNB setzt strategisch auf RDA und überarbeitet ihr Geschäftsmodell für Datendienste dahingehend, als dass “intensivere Kooperationen und eine größere Präsenz bibliothekarischer Informationen im Internet der Daten“ angestrebt werden sollen.[197] Die Library of Congress hat mit der Veröffentlichung der LCSH in SKOS bereits einen konkreten bibliothekari-

[192] [Yang et al. 2009], 19ff
[193] Vgl. [Pohl 2010]
[194] Vgl. http://www.daveyp.com/blog/archives/528
[195] http://www.hbz-nrw.de/dokumentencenter/presse/pm/datenfreigabe
[196] http://www.flickr.com/photos/danbri/4326955233/
[197] [Kett 2009], 9

schen Beitrag offener Daten im Web geliefert. Daneben wurden mit dem Projekt *LCCN Permalink* stabile und einheitliche URI-Identifier geschaffen, um Entitäten und ihre Datensätze konkret identifizieren zu können.[198] Die Ausführungen bei [Marcum 2008] in Reaktion auf den Bericht der *Working Group on the Future of Bibliographic Control* lassen vermuten, dass weitere Konversionen kontrollierter Vokabulare geplant sind.[199]

5.2 Implikationen und Forschungsfelder

Durch die Eingliederung bestehender bibliographischer Daten in ein einheitliches Datenmodell wie RDF ist mit einer Vielzahl auftretender Inkonsistenzen zu rechnen, die gegenwärtig in den verwendeten Datenstrukturen verborgen sind. Ihre Auflösung und Integration ist ein weites zukünftiges Forschungsfeld, zu dem [Yee 2009] und [Styles et al. 2008] erste Überlegungen angestellt haben. Aus der Perspektive der Katalogisierung verfasst und daher skeptischer Natur, befasst sich der Artikel von Yee mit den Möglichkeiten einer Wiedergabe der Dimension gegenwärtiger Erschließung in Verbindung mit der Zuweisung von FRBR-Entitäten im RDF-Modell. Dabei werden wichtige Aspekte wie die Eindeutigkeit von FRBR-Entitäten thematisiert und Yees Auseinandersetzung bietet eine gelungene Diskussionsgrundlage für die fachliche Debatte. Obwohl die Notwendigkeit eines neuen Datenmodells nicht bestritten wird, bleibt Yee in ihren Ausführungen dennoch teilweise den Paradigmen traditioneller Metadatenverarbeitung verhaftet. Danach sei die Realisierung der Anzeige von bibliographischen Daten die Hauptaufgabe der Katalogisierung und eine angestrebte Granularitätstiefe von Metadaten an diesem Ziel auszurichten. Linked Data ist jedoch viel mehr als der Versuch einer Replikation bestehender Datenstrukturen und ihrer Abbildung für Endnutzer, wie auch Yee ahnt:

> *"To a cataloger, it looks as though the plan is for RDF data to float around loose without any requirement that there be a method for pulling it together into coherent displays designed for human beings."* [200]

Daraus ergibt sich die wichtigste Implikation einer bibliothekarischen Adaptierung von Linked Data: die Auflösung des gängigen Datensatzbegriffs. Angesichts einer Informationsinfrastruktur, in welcher Daten nahtlos ineinander

[198] Vgl. z.B. http://lccn.loc.gov/66012340
[199] Vgl. [Marcum 2008], 43ff
[200] [Yee 2009], 67