Linked Open Projects: Nachnutzung von Projektergebnissen als Linked Data

Kai Eckert
Universitätsbibliothek Mannheim

Semantic Web & Linked Data - Elemente zukünftiger Informationsinfrastrukturen,
1. DGI Konferenz, 62. Jahrestagung der DGI, Frankfurt am Main, 7. bis 9. Oktober 2010

Preprint

Abstract

In vielen wissenschaftlichen Projekten – nicht nur im Bibliotheksbereich – geht es um die Erzeugung von Daten, häufig mit Hilfe automatischer Verfahren. Die Nachnutzung dieser Daten gestaltet sich häufig schwierig. In diesem Artikel werden wissenschaftliche Projekte beschrieben, die an der Universitätsbibliothek Mannheim durchgeführt wurden und werden. Anhand einfacher Beispiele wird gezeigt, wie durch Linked Data die Daten, die in diesen Projekten generiert werden, leicht und und flexibel nachgenutzt werden können.

Einleitung

Die Nachnutzbarkeit von Ergebnissen spielt eine große Rolle bei der Förderung wissenschaftlicher Projekte, insbesondere, wenn sie zum Ziel haben, neue Verfahren in einer Anwendungsdomäne zum Einsatz zu bringen.

Es liegt zwar in der Natur der Sache, dass oft im Vorfeld nicht genau klar ist, was das Ergebnis eines wissenschaftlichen Projekts sein wird; aber trotzdem spielt schon bei der Beantragung von Fördermitteln die potentielle Nachnutzung eine wichtige Rolle. Wie genau eine solche Nachnutzung aussieht, ist dabei sehr unterschiedlich und von der Art des Forschungsvorhabens abhängig.

Im Rahmen dieses Artikels werden vor allem anwendungsorientierte Vorhaben aus der Informationswissenschaft betrachtet. Bei solchen Projekten geht es häufig um die Entwicklung von Systemen, die Daten mittels neuartiger Verfahren erzeugen oder vorhandene Daten aufbereiten. Zwei solche Projekte, die an der Universitätsbibliothek Mannheim durchgeführt wurden, dienen im Folgenden als Beispiel:

1. Ein System zur automatischen Klassifikation von Veröffentlichungen basierend auf Titeldaten und
2. ein System zur automatischen inhaltlichen Erschließung basierend auf Abstracts.

Die grundsätzliche Nachnutzbarkeit der erzeugten Daten steht in beiden Fällen außer Frage; interessanter ist, wie man diese Daten idealerweise bereitstellt, um eine einfache und flexible Nachnutzung zu ermöglichen. Eine vielversprechende Variante dafür ist Linked Data, eine Technologie, deren Hintergrund und mögliche Anwendung ebenfalls anhand einiger Beispiele vorgestellt werden.

Projekte der UB Mannheim

An der Universitätsbibliothek Mannheim gab und gibt es verschiedene Projekte, die sich mit der automatischen Verarbeitung und Anreicherung von bibliographischen Informationen befassen.

Das erste Projekt hat das Ziel, Publikationen auf der Basis von Titeldaten automatisch zu

klassifizieren. Bei den Arbeiten zu diesem Projekt hat sich der Bedarf für eine Vorverarbeitung der Titeldaten ergeben, die ihrerseits interessant genug für andere Nachnutzungsmöglichkeiten sind. Insofern geht es eigentlich eher um drei Projekte:

Automatische Klassifikation. Das Projekt zur automatischen Klassifikation läuft seit 2005. Es geht dabei darum, Klassifikationen aus der Regensburger Verbundklassifikation automatisch einem neuen Titel zuzuweisen. Dazu werden ähnliche Titel aus der vorhandenen Datenbasis (SWB und HeBIS Katalogdaten, > 22,6 Millionen Einträge) gesucht und mit diesen verknüpfte Klassifikationen gewichtet und als mögliche Kandidaten dem neuen Titel zugewiesen. Ziel ist es, jedem Titel mindestens eine RVK-Klassifikation zuzuweisen.

Konkordanz HeBIS-SWD. Zur Unterstützung der automatischen Klassifikation ist es hilfreich, Titel, die unter klassifikatorischen Aspekten als gleichwertig anzusehen sind, zunächst zusammenzufassen. Dieses Teilprojekt läuft seit 2010 und das Ziel ist es, gleiche Titel, die in beiden Verbundkatalogen vorkommen – aber auch verschiedene Auflagen des gleichen Werkes – zusammenzuführen. Auch dabei kommt ein automatisches Verfahren zum Einsatz.

Konkordanz RVK-SWD. Ebenfalls als Unterstützung wird seit 2009 die Korrelation zwischen den vergebenen Schlagworten aus der SWD und der zugeordneten Klassifikation nach RVK untersucht.

Das zweite Projekt befasst sich mit der automatischen Inhaltserschließung von wissenschaftlichen Abstracts nach einem vorgegebenen Thesaurus:

Automatische Inhaltserschließung. Ziel dieses Projekts ist es, Begriffe aus dem Standard-Thesaurus Wirtschaft (STW) gegebenen Abstracts von wissenschaftlichen Veröffentlichungen zuzuweisen. Dabei kommt eine kommerzielle thesaurusbasierte Suchmaschine zum Einsatz. Das Projekt lief von 2007 – 2009, unter anderem wurden dabei rund 10.000 Artikel aus den Springer-Nationallizenzen erschlossen.

Allen hier vorgestellten Projekten und Teilprojekten ist gemeinsam, dass als Ergebnis Daten produziert werden, die für sich genommen interessant für eine weitere Nachnutzung sind. Konkret werden die folgenden Daten erzeugt:

1. Mögliche **RVK-Klassifikation** für einen gegebenen Titel in SWB oder HeBIS, der noch keine Klassifikation hat.
2. **Titel** aus SWB und HeBIS, die bezüglich ihrer Klassifikation als gleichwertig anzusehen sind für einen gegebenen Titel in SWB oder HeBIS.
3. Die **Korrelation zu Schlagworten** aus der SWD, bzw. Notationen aus der RVK für eine gegebene RVK-Notation, bzw. ein gegebenes SWD-Schlagwort.
4. **Begriffe aus dem STW** für ein gegebenes Abstract.

Nachnutzung der erzeugten Daten

Alle vorgestellten Verfahren können und wollen einen Bibliothekar nicht ersetzen. Gerade was die inhaltliche Erschließung angeht, erreichen automatische Verfahren im Allgemeinen nicht die geforderte Qualität an eine intellektuelle Erschließung. Dadurch ist eine Nachnutzung dieser Daten in den vorhandenen Systemen – zum Beispiel das Einspielen in einen Verbundkatalog – schwierig, wenn man eine Kompromittierung der Qualität der vorhandenen Daten vermeiden möchte. Zumindest würde man eine weitere Qualifizierung der Daten benötigen, um sie als automatisch generiert zu kennzeichnen.

Trotzdem sind die Einsatzmöglichkeiten der Daten vielfältig und die UB Mannheim ist bestrebt, die Daten Interessenten in geeigneter Form zur Verfügung zu stellen. Eine gängige Form dafür ist die Bereitstellung der Daten als einfache Komma-separierte Textdatei (CSV) oder als XML. Die folgende Tabelle listet einige Kriterien und Anforderungen für die Bereitstellung von Daten auf und bewertet kurz die Bereitstellung als CSV/XML:

Tabelle 1: Kriterien zur Datenaufbereitung und Darstellung CSV/XML

	Anforderungen	CSV/XML	Bewertung
Datenformat	Einheitliche Syntax	CSV, XML	Maschinenlesbar, aber nicht maschinenverstehbar
	Maschinenlesbar		
Ressourcen	Verknüpfung mit den Ursprungsdaten	Identnummer	Oft nur im gegebenen Kontext eindeutig
Vokabular	Nachvollziehbare Semantik	Spaltenüberschrift, Feldbeschreibungen	Uneinheitlich, Dokumentation zum Verständnis der Daten erforderlich
	Projektübergreifend	Bestenfalls innerhalb einer Institution	
Zugang	Einheitlicher Zugriff	Archive im Web	Unhandlich
	Möglichkeit der Selektion	Archivebene	Allenfalls sehr grobe Untergliederung

Kurz gesagt: Das Problem bei der Bereitstellung von Daten in dieser Form ist, dass von einem Interessenten immer ein beträchtlicher Aufwand betrieben werden muss, um die Daten zu nutzen. Auch wenn die Daten frei zugänglich und maschinenlesbar sind, muss erst verstanden werden, welche Daten hier für welche Ressourcen bereitgestellt werden. Häufig werden die Ressourcen innerhalb der eigenen Anwendung anders identifiziert und es muss erst eine Zuordnung erfolgen. Es ist kaum möglich, nur auf einzelne Datensätze zuzugreifen oder auch nur herauszufinden, ob für eigene Ressourcen Daten vorliegen, die man nutzen könnte.

Linked Data

Für all diese Probleme bietet Linked Data – also die Bereitstellung der Daten mit Hilfe des Ressource Description Frameworks (RDF) – eine elegante Lösung. Dabei unterscheidet sich diese Lösung gar nicht so gravierend von der Bereitstellung als CSV oder XML. Letztlich ist RDF auch „nur“ ein Austauschformat für Daten. Die folgende Tabelle stellt Linked Data der gängigen Praxis gegenüber:

Tabelle 2: Vergleich CSV/XML mit RDF

	Anforderungen	CSV/XML	Linked Data
Datenformat	Einheitliche Syntax	CSV, XML	RDF
	Maschinenlesbar		
Ressourcen	Verknüpfung mit den Ursprungsdaten	Identnummer	URI
Vokabular	Nachvollziehbare Semantik	Spaltenüberschrift, Feldbeschreibungen	Einheitliches Vokabular
	Projektübergreifend	Bestenfalls innerhalb einer Institution	
Zugang	Einheitlicher Zugriff	Archive im Web	URI per HTTP

	Anforderungen	CSV/XML	Linked Data
	Möglichkeit der Selektion	Archivebene	Abfrage über SPARQL

Welche Vorteile bietet Linked Data also konkret gegenüber CSV/XML?

Einheitliches Datenmodell. Der erste Vorteil liegt in dem sehr einfachen und einheitlichen Datenmodell, das durch RDF vorgegeben wird: Alle Daten werden als Aussagen formuliert, die jeweils durch ein Tripel – bestehend aus Subjekt, Prädikat und Objetk – repräsentiert werden. Ein Beispiel für eine solche Aussage ist:

```
<Katalogeintrag 234> <gehört zu> <Klassifikation 765>.
```

Durch diese Einheitlichkeit ist schon viel gewonnen. Selbst wenn der interessierte Nachnutzer nicht weiß, was mit `<gehört zu>` und `<Klassifikation 765>` gemeint ist, weiß er trotzdem, dass hier eine Aussage über `<Katalogeintrag 234>` gemacht wird, also Zusatzinformationen zu diesem zur Verfügung stehen.

Einheitliche Identifikatoren: URIs. Die Frage, die sich aus dem vorherigen Beispiel ergibt, ist: Was genau ist `<Katalogeintrag 234>`? In RDF werden sowohl die Subjekte, als auch die Prädikate durch sogenannte Uniform Resource Identifiers (URIs) weltweit eindeutig und dauerhaft identifiziert. Dabei kann es durchaus mehrere URIs für die gleiche Ressource geben, etwa weil jeder Verbund eine eigene URI vergibt für ein neues Katalogisat.

Aber eine URI kann nie mehr als eine Ressource bezeichnen und nachdem einmal eine Ressource mit einer URI versehen wurde, kann diese URI nicht mehr zur Identifikation einer anderen Ressource verwendet werden (Persistenz).

URIs haben üblicherweise die Form von Webadressen (URLs), was viele Vorteile mit sich bringt. Zum einen kann man die URI als URL interpretieren und als Anbieter unter dieser URL Daten zu der fraglichen Ressource bereitstellen. Zum anderen entsteht dadurch die gewünschte weltweite Eindeutigkeit, indem jeder Anbieter von Daten nur URIs vergibt, die mit einer zum Anbieter gehörenden Domain beginnen.

An der UB Mannheim werden einerseits URIs genutzt, um die Titeldaten in den Verbünden eindeutig zu identifizieren (z.B. http://data.bib.uni-mannheim.de/data/swb/002773333), andererseits, um das Vokabular zur Bereitstellung der Projektergebnisse zu definieren (z.B. http://data.bib.uni-mannheim.de/data/terms/equalsForClassification).

Verknüpfbarkeit von Daten und Vokabularien. Damit aus Daten verknüpfte Daten – Linked Data – werden, müssen diese verknüpfbar sein. Die Verknüpfung erfolgt entweder direkt, indem man mittels RDF Ressourcen mit anderen Ressourcen in Bezug setzt, z.B. durch die Zuweisung eines Autors über die von der Nationalbibliothek vorgegebene URI für Personen aus der Personennormdatei (z.B. Bertolt Brecht: http://d-nb.info/gnd/118514768).

Alternativ kann eine Zuordnung auch implizit erfolgen, indem man eine Äquivalenz zwischen einer lokal verwendeten URI und einer anderen URI erklärt; etwa, weil beide den gleichen Autor bezeichnen. Das kann insbesondere auch für die URIs zur Definition der Prädikate erfolgen, wodurch sich leicht Aussagen von anderen zu den eigenen Aussagen über Ressourcen in Bezug setzen lassen.

Zugang. Zuletzt bietet Linked Data auch verschiedene, sich gut ergänzende Möglichkeiten, auf bereitgestellte Daten zuzugreifen. Im einfachsten Fall kann man die RDF-Daten auch weiterhin komplett in Archiven zum Download bereitstellen. Das ist natürlich relativ unhandlich, aber in

vielen Fällen trotzdem wünschenswert, etwa wenn man große Datenbestände, die sich nur selten ändern, komplett übernehmen und in sein lokales System integrieren möchte.

Daneben besteht die Möglichkeit, die eigenen Daten nur mit den Fremddaten zu verknüpfen und dadurch lediglich im Bedarfsfall die zusätzlich hinterlegten Informationen vom Bereitsteller der Fremddaten abzurufen.

Die folgende Abbildung zeigt eine Webseite aus dem Linked Data Service der UB Mannheim, auf der Informationen zu einer URI (Katalogisat aus dem SWB, http://data.bib.uni-mannheim.de/data/swb/106344897) menschenlesbar aufbereitet werden. Links zu anderen Ressourcen werden dargestellt und beim Anklicken werden die Daten, die vom Fremdanbieter bereitgestellt werden, geladen und dem Benutzer angezeigt:

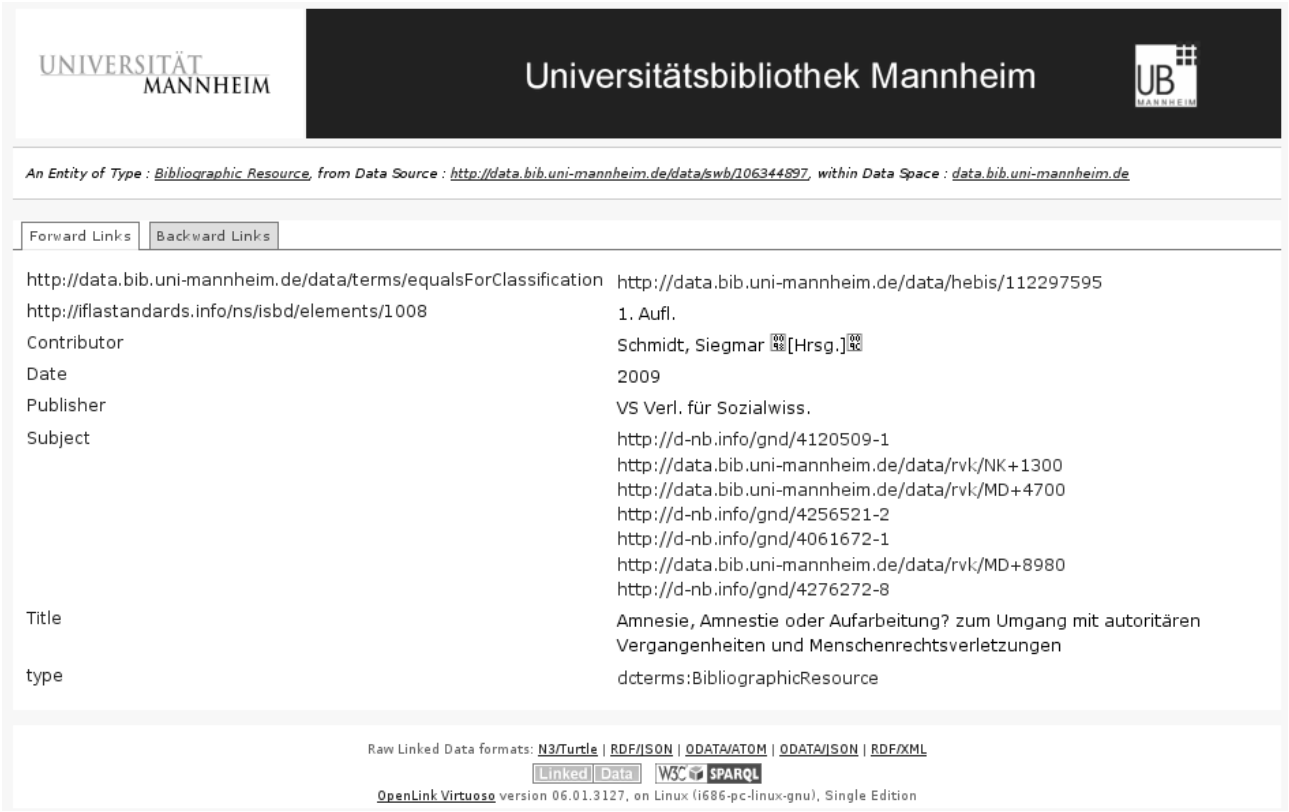

Abbildung 1: Linked Data Service der UB Mannheim

Die dritte und für im Rahmen der Nachnutzung von Projektergebnissen spannendste Form ist aber der Zugang zu den bereitgestellten Daten über eine SPARQL-Schnittstelle. SPARQL ist eine Abfragesprache für RDF und erlaubt es, gezielt nach Aussagen über eine gegebene Ressource zu recherchieren. Da es für alle gängigen Programmiersprachen Frameworks für RDF und SPARQL gibt, lassen sich die Daten damit genau so leicht abfragen, wie mit SQL in einer lokalen Datenbank.

Möchte man also auf die Ergebnisse des Projekts zugreifen, bei dem Konkordanzen zwischen HeBIS und SWB erstellt wurden, so kann über die SPARQL Schnittstelle der UB Mannheim (http://data.bib.uni-mannheim.de/sparql) die folgende Anfrage gestellt werden:

```
select distinct ?URI where {<http://data.bib.uni-mannheim.de/data/swb/107126095> <http://data.bib.uni-mannheim.de/data/terms/equalsForClassification> ?URI}
```

In natürlicher Sprache liest sich das als: „Gib mir alle URIs für Ressourcen, die bezüglich ihrer Klassifikation als gleichwertig zu dem Katalogisat mit der PPN 107126095 im Verbundkatalog des SWB angesehen werden.“ Als Ergebnis erhält man die URIs für alle Ressourcen, die aktuell durch das an der Universität Mannheim entwickelte automatische Verfahren als gleich ermittelt wurden:

http://data.bib.uni-mannheim.de/data/swb/302937129 und http://data.bib.uni-mannheim.de/data/swb/284337498.

Zusammenfassung

In diesem Artikel wurde anhand einiger Beispiele gezeigt, wie sich Daten, die als Ergebnisse innerhalb von Projekten erzeugt werden, als Linked Data darstellen lassen. So lassen sich die Daten deutlich leichter nachnutzen, als als Rohdaten in einem maschinenlesbaren Format, weil weniger Dokumentation bezüglich des Datenformats benötigt wird und weil die Integration von Daten aus unterschiedlichen Quellen von RDF direkt unterstützt wird.

Linked Data ist dabei keine komplette Neuerfindung des Datenaustauschs und löst sicher auch nicht alle Probleme, die sich bei der Integration von Fremddaten in eigene Systeme ergeben, aber RDF als technische Basis von Linked Data bietet viele Verbesserungen im Detail gegenüber CSV oder XML, die es tatsächlich zu einer idealen Plattform für die Publikation von Projektergebnissen machen.

Linked Data kann sein volles Potential nur entfalten, wenn möglichst viele Daten als Linked Data zur Verfügung stehen. Im Umfeld der Bibliotheken wäre es daher wünschenswert, dass die existierenden Daten – Titeldaten und Normdaten – als Linked Data publiziert werden oder zumindest die Datensätze einheitlich über URIs identifizierbar gemacht werden. Erst dann ist es ein leichtes, beliebige neue Informationen, wie sie in wissenschaftlichen Projekten durch automatische Verfahren erstellt werden, mit den vorhandenen Daten zu verknüpfen und sie bei Bedarf in neuen Anwendungen auch tatsächlich nachzunutzen.