Helen Williams

Linked data and libraries

Article (Accepted version) (Unrefereed)

Talis Open Day: Linked Data and Libraries

Helen Williams , London School of Economics

'Linked Data and Libraries' was organised by Talis at the British Library on 21st July 2010 with the aim of introducing participants to the principles, practice and potential of Linked Data.

After a welcome and introductions by Richard Wallis of Talis, Zach Beauvais, editor of Talis' *Nodalities* magazine, gave a brief introduction to Talis and the world of Linked Data. Talis is involved in building next generation web applications and the purpose of *Nodalities* is to 'bridge the divide between those building the Semantic Web and those interesting in applying it to their business requirements'.
For further details see http://www.talis.com/nodalities/ Early issues of *Nodalities* addressed the possibilities offered by the Semantic Web while more recent issues contain articles which detail work in progress, reflecting the way in which the web of data has been emerging over the last two or three years.

The second speaker, Romain Wenz from the Bibliothèque Nationale de France (BNF), talked about their Linked Data work; the data bnf.fr project. This project is seeking to make library resources easier to find on the web by creating web pages which group resources round the concepts of works and authors. Using HTML to build the web pages the data will be gathered from both internal and external sources to create an information hub which can be browsed both by users and search engines. The library currently has eight different catalogues which will need to be internally interoperable for this to work so the project will require significant changes to library policies in order for it to be a success. As well as addressing issues of efficiency and interoperability the project needs to be viable in the Semantic Web environment. This means that data needs to be accessible in RDF (Resource Description Framework), and that URIs (Uniform Resource Identifiers) need to be maintained and to remain stable. The project is currently at the early stages, but will ultimately make retrieved data more valuable for users.

Once we had seen how the BNF are hoping to harness the possibilities of Linked Data, Rob Styles from Talis took us back to basics on Linked Data, RDF and SPARQL, with a clear presentation on these technical topics. He explained Linked Data as the technical practices employed to create a web of data. Today the web is a global network of linked HTML documents. As a place where anyone can say anything about anything it is a decentralised and unregulated environment. The documents making up this vast collection of 'human readable' knowledge and opinion are linked, but the links are not qualified and so there is no meaning in them.

The web of data is 'socially' still the same decentralised, unregulated environment, but in contrast to the web we are familiar with, it is a global network of linked statements. The links between these statements are qualified, which means that the data is machine readable instead of only 'human readable'.

The technology underpinning the web of data is RDF (Resource Description Framework), which Rob described as a simple and powerful data model specifically designed to work with the architecture of the web and define

relationships between things. It means that 'human readable' web page statements can be translated into machine readable statements and therefore can be analysed, searched and indexed in information retrieval. In RDF the resources involved, such as web pages, people or places, are named using URIs. Two people using the same URI are assumed to be describing the same thing and so when these are merged, they become one URI containing the properties of both datasets. Using data from many different websites this can create a vast amount of Linked Data. Rob pointed out that one does, of course, need to trust the sources from which the data is obtained, or to have some kind of internal validation process.

The final part of Rob's talk dealt with some more technical matters, covering the ways in which RDF can be written down using Turtle, N-triples, RDF-XML, RDF-JSON and RDFa, before moving onto SPARQL (SPARQL Protocol and RDF Query Language). This is a graph-pattern-based query language for RDF which allows searches to be carried out using the queries 'Ask, Select, Describe and Construct'.

In the final session before lunch we turned from theory to practice as Richard Wallis showed us some sites already using Linked Data together with the RDF data behind them. To pick up on just a couple of these examples, we looked at the catalogue of the Royal Scottish Academy of Music and Drama http://prism.talis.com/rsamd/ where we saw a standard catalogue interface with results and facets. What users do not see is that this is running on top of a Linked Data dataset which means that results pages can link out to other related items. In this case the links are to other items in the same data set, but a site such as BBC wildlife http://www.bbc.co.uk/nature/animals/ has taken this a step further. Building their web pages on Linked Data principles and using URIs as identifiers, they are able to pull in data from other well known and reliable sources rather than gathering and holding all the information themselves.

After lunch we heard 'Lightening Talks' from three of the Open Day attendees. First was Neil Wilson, Head of Bibliographic Development at the British Library, talking about plans for using BL data in new ways. He referred to the Government's 2009 commitment to the principle of opening up public data and facilitating greater use of Linked Data. Neil pointed out that an open data strategy has certain financial and legal implications such as loss of revenue from traditional bibliographic services and increasingly complex copyright issues, and also reputational implications in terms of whether the library is perceived to be 'keeping up'.

The BL is looking to go beyond traditional MARC data and has already developed the capability to support non-library metadata standards as used by the wider web community. In May 2010 they hosted a Linked Data workshop to address technical challenges, standards and prototypes, and they are now doing some further work with JISC and participating in an Open Bibliography project with Cambridge University Library.

Next, Sally Chambers from the European Library spoke about the Europeana Data Model. The European Library already offers access to the resources of 49 libraries in Europe and Europeana's aim is to make digital information from libraries, museums, archives and audio-visual collections within Europe accessible. Using the

European Data Model, which has been produced to facilitate participation in the Semantic Web, Europeana is working on a framework for connecting, collecting and enriching metadata to enable the browsing of these digital objects in new ways.

Finally Felix Ostrowski from the Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen spoke about building a Linked Data based index of library institutions. Currently the data held in the various sources which might inform such an index is only very basic or is outdated. Up to date or detailed information is available on library websites, but these are usually only 'human readable'. There are still decisions to be made surrounding copyright issues and modelling decisions, but if libraries can be encouraged to use RDFa in their web pages, then their data is turned into a machine readable interface and the quality of the data which could be harvested to a centralised source is greatly enhanced.

Rob Styles followed the 'Lightening Talks' with a presentation showing us how MARC data can be transformed into RDF. Each RDF property (for example, author, title, subject) is labeled and given an http URI so that it can be an RDF link on the Web. Authority and bibliographic data, which are kept separately in MARC, are meshed together in RDF through sharing URIs, thereby bringing data together. The aim is to create, in the web, models of data which identify all the things the MARC record describes but using different technology. For librarians some of the important developments are Library of Congress Subject Headings as Linked Data http://id.loc.gov/, the Virtual International Authority File (VIAF) http://viaf.org/, and the work in progress to publish Resource Description and Access (RDA) vocabulary as RDF http://metadataregistry.org/.

Antoine Isaac spoke about the W3C Library Linked Data Incubator Group. This is a group which is running for one year, not to make recommendations about Linked Data, but rather to look at innovative ideas in this area. Antoine highlighted that though libraries have a long history in producing metadata, usually of a high quality, this metadata often remains locked in bibliographic records. This group hopes to help increase the global interoperability of library data on the web by bringing together people who are involved in the Semantic Web, building on existing initiatives and identifying areas of collaboration for the future. As well as gathering case studies of success stories the group will identify relevant data models, vocabularies and ontologies and consider the need for new standards, guidelines and best practices.

Before an open discussion to end the day Richard Wallis gave an overview of the Talis platform, showcasing the cloud-based data storage on offer from Talis as a Linked Data publishing platform.

The day was very informative and it was exciting to hear about all the Linked Data projects that are already in progress and to consider the extent of future possibilities for libraries.

Videos from the day can be seen at:

http://blogs.talis.com/nodalities/2010/08/linked-data-and-libraries-almost-like-being-there.php

This article is based on the presentations given at the Talis Open Day at the British Library on 21st July 2010 and credit should be attributed to Richard Wallis, Zach Beauvais, Romain Wenz, Rob Styles, Neil Wilson, Sally Chambers, Felix Ostrowski and Antoine Isaac.