Article on Richard Wallis talk @ Ebooks 2013 conference, UCL. Richard Wallis (formerly Talis; now Tech. evangelist @ OCLC and consultant) – "Linked Data for Ebook Discovery"

Gopal Dutta, Information Services Librarian, Leeds Metropolitan University

This article is a write-up of a talk I attended in May 2013 at the "Ebooks 2013" conference organised by the UCL DIS team.

Richard Wallis' talk resonated with me particularly as I've a dual-role as Information Services Librarian at LeedsMet University: cataloguing and ebooks management.

Wallis began his talk by explaining that very few students begin their information journeys with library public catalogues or discovery systems. Predictably, most queries begin on Google, where "Google" is used as a catch-all term to cover all major search engines: Bing, Yahoo, Baidu etc.

However, even Google is now having trouble tracking down resources, due to the wealth of information. Searching for ebooks on Google, the first page of results are more likely to be secondary sources of information, for example, reviews, than the item itself. So, Google have started to use the white space on the right of the results page to display "knowledge graphs" which attempt to interpret the search query and understand that the user is looking for a certain "thing" rather than an endless series of "strings" (http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html).

Richard suggested that libraries need to start doing something similar, in order to best facilitate the discovery of their (expensively purchased) resources. The reasoning goes: if students are using Google to look for the ebooks that we are buying for them, should we not ensure that they are as discoverable as possible? He thus introduced www.schema.org, a "broad and shallow" cataloguing vocabulary, which can be seen in OCLC / WorldCat records. You can see what a schema record looks like by scrolling to the bottom of any OCLC / WorldCat catalogue record and opening the "linked data" tab.

This shows the metadata for the records in HTML format, rather than RDA or AACR2. The HTML allows the data to be structured and therefore extensible across the open web – linked data.

Wallis was careful to explain the limitation: "...schema.org is a very broad, fairly shallow vocabulary, which in no way can be seen as a replacement for traditional library catalogue vocabularies."

The actual vocabulary has been established by the large search engine companies (Google et al), using the RDF standard as an underlying language. Because it is simple to use, it is one of the most popular structured data languages in use on the World Wide Web.

Wallis argues that it should not be seen as a replacement for traditional cataloguing in libraries, as it does not have the deep, descriptive capability of RDA or AACR2. Nevertheless, it has become a standard because of this shallowness and ease of use. It carries enough information for it to be useful. For example, it can carry information explaining that the object in question is a book. It can also automatically link across different interfaces.

So why should cataloguers care about this?

Cataloguers traditionally describe information objects at the "manifestation" level. We are trained to "catalogue the item in front of you."

However, the rest of the world search for items at the "work" level i.e. people don't really care about the format anymore. As e-reading capabilities increase, people will not care so much whether the object is printed – full text access will be the key issue for people searching. This connects to the recent changes that are happening with RDA and the 3XX fields, which will start to carry "manifestation" information.

So what do cataloguers need to do?

Wallis advises that cataloguers “stop copying and start linking” and introduced the term “catalinking” as a new way of thinking about cataloguing. Eric Miller is credited as the originator of the term catalinking (http://zepheira.com/about/people/eric-miller/)

The suggestion for cataloguers is that rather than searching authority files for the correct data and then copying the text into local catalogues, they should instead link to persistent URIs of the same authorities. This then allows their catalogue to be part of the linked web of data. In practical terms, this could work with the linked data appearing as a widget or add-on to a traditional bibliographic record.

I was left slightly unsure as to next steps. On the one hand, an add-on which incorporated the schema.org fields into a traditional cataloguing record does not seem such a difficult leap to make. On the other, it would require a great deal of cooperation between different libraries before they could agree on the fundamental facets of the new standard – it doesn’t seem quite right that major libraries would agree to use schema.org simply because Google et al are using it. In a sense, Wallis was arguing for something not too distinct from RDA, except a quicker, more practical solution. So perhaps the answer lies somewhere between the two?

I’m also not sure whether linked data would accomplish the task of pushing library catalogue records to the top of searches across the big search engines. For this to happen, I suspect that entire cataloguing web interfaces would require rewriting, in a way that makes them much more “open” and amenable to web crawlers.

Despite these reservations, I found Wallis’ talk very thought provoking. “Catalinking” is an intriguing concept and further developments have just been announced, with OCLC releasing several million “work” descriptions (http://semanticweb.com/194-million-linked-open-data-bibliographic-work-descriptions-released-oclc_b41921)

Richard Wallis’ talk from Ebooks 2013 is available to view in full here http://river-valley.tv/use-of-linked-data-for-ebook-discovery/ and you can follow him on twitter here: https://twitter.com/rjw