Report on the CIG event: Linked Data – what cataloguers need to know, held 20.02.2015

Article (Accepted Version)

Clare Playforth

Report on the CIG event: ' Linked Data – What Cataloguers need to know - 20.02.2015

Speakers: Tom Meehan (UCL), Owen Stephens (Owen Stephens Consulting), Corine Deliot (British Library), Alan Danskin (British Library).

The day began with an Introduction to Linked Data from Tom Meehan. He explained that we can't just go ahead and start cataloguing Linked Data but MARC is on its way out and Linked Data will most likely replace it possibly (although not necessarily) in the shape of BIBFRAME. He then went on to talk about MARC data and described how it is structured and labelled in a recognised format so that we (libraries) can share it with each other. We were also reminded that MARC records are for the most part not open and that people often don't have explicit sharing licences in place. Tom then demonstrated what Linked Data looks like when it isn't 'open' by showing us a blank slide. There is no point in creating Linked Data unless you make it open and by open we mean freely available and licensed to be re-used, redistributed, and repurposed. To provide our definition of Linked Open Data we were presented with Tim Berners-Lee's guidelines…

- use HTTP URIs as names for things.
- put it on the web and make it open.
- when someone looks up a URI provide useful information using the standards RDF (Resource Description Framework) and SPARQL (**S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage).
- include links to other URIs so that people can discover more things.

This gave us a broad understanding of what Linked Data means and next we learnt how to construct an RDF triple composed of a subject, predicate and object using URIs. We were shown how to write it in Turtle, which is easier to read, and how to use prefixes for our chosen vocabularies to save ourselves some typing. There are lots of ways to combine the various vocabularies and you can choose any combination you like to model the data – whichever is most suited to your library's need. There is also the option of making up your own schema should existing ones

not be appropriate. Tom described the RDF and Linked Data in a way that I felt most people attending would have understood – we were taken through the processes step by step and it all seemed pretty clear to me. I did feel like I'd had a bit of a head start as I'd written an essay on Linked Data for my MSc the previous month but I think as an introduction to the topic this couldn't have been better. We were able to view a 'real life' example of Linked Data through an OCLC WorldCat record. OCLC use schema.org and RDFa for their model and we were shown the Linked Data tab at the bottom of the page which allows you to view the N triples or Turtle. I think most people will have found this useful as it gave us a way to study the data after the event and to try to make sense of what we have learned.

Next up was Owen Stephens to talk to us about how to publish and use Linked Data. He described how to create the URIs themselves and how to make them cool! URIs are fundamental to Linked Data because these *are* the links. He explained how it's important to be very precise in what you are identifying, for example London the 'thing' and London the 'concept' are two different statements in RDF. Also we need to think about page URIs. Should we re-direct the link to the page or to what the page is about? Apparently there is a lot of ambiguity and a lot of time spent in the Linked Data community arguing about it. With library data the URIs need to not change or disappear. It's good practice therefore to use Cool URIs which should have a numerical unique identifier and avoid things subject to change (such as terms, names or programming languages). I think this would have struck a chord with the audience and I expect everyone realised at this point how traditional library authority files are not accurate enough when it comes to identifying things uniquely.

Following this Owen gave us a rundown of the vocabularies and ontologies we might consider using before explaining the different publishing models:

- Static files of RDF statements stored like you would do HTML pages on a web server (not always appropriate for a whole catalogue).
- Dynamically generated views where you have data in the back end of the system and then generate views of it when people want to look at it i.e. when the page is viewed. Most RDF is published this way.

- Embedded Linked Data within HTML documents. This can be done with existing catalogues and it's what OCLC have done with WorldCat. Generating these views is not necessarily a difficult step but it does involve work for systems and cataloguers.

As we had already learnt from Tom this morning there is no point creating Linked Data unless you are going to make it open so Owen briefly discussed what this might entail in terms of licenses and copyright. We need to keep track of licensing across all the data from multiple triple sources, and this may not be easy but it is certainly possible and metadata licensing is not a new thing. Current Linked Data models have used various Creative Commons or Public Domain licenses.

Corine Deliot from the British Library was next to present, describing how they went about publishing the British National Biography (BNB) as Linked Open Data. It was really great to have an actual use case described to us and she gave us a nicely detailed description of the project and the workflows. She began by giving the justification for carrying out the project in the first place saying that since 2009 the UK government has been committed to public bodies having open data and the British Library particularly wanted to look beyond library formats and start trying to adapt to cross domain standards. She said that they tried to be pragmatic and choose a definite data set to work with which is why they went with the BNB. This record set also benefits from data that is as consistent as can be reasonably expected. I was interested to learn that they used existing staff and resources and built on their existing skills with MARC. They had no programmers or data architects involved but they did get training and mentoring from external provider, Talis who also hosted the data for them afterwards.

Next she described the process they used to model the data and explained how they went through the various steps highlighted previously by Tom and Owen. We were shown the BNB data model slide and its complexity drew a few gasps from the room but she assured us that it is not as complicated as it looks and that models such as these can be useful tools to refer to when creating SPARQL queries. In terms of

usage statistics Corine told us that they were still gathering data and that due to the necessary 'openness' it will always be hard to find out who is using it. They are currently gathering information on the number of hits on the SPARQL endpoint, the number of downloads from the British Library webpage, weblogs and analysis reports and also anecdotally from tweets etc.

We were treated to a couple of examples where they had to find their own way of doing things (specifically modelling publication as an event). Overall it is clear to see that if we were creating Linked Data from scratch it would be easy but it's the transforming of legacy data (MARC records in the majority of cases) that makes it hard. This reminds me how valuable it is not just to have open Linked Data but to be open with the projects and procedures used to create it. If people are willing, like Corine and the British Library, to share their work with us then we can learn from their processes and benefit from a collective knowledge on the topic. Indeed towards the end of her talk she suggested it would be good to get some more collaboration with other libraries.

The following presentation was the one I had been looking forward to the most. We got to get our hands dirty having a go at querying the data in the BNB with SPARQL in a session led by Owen. After a brief introduction to the RDF query language we followed Owen and wrote queries in the BNB SPARQL editor. I found this fun and, although I got a bit lost towards the end I was able to get some of the SELECT queries to return results. At this basic level the syntax was similar enough to SQL for me to feel comfortable but the speed of the session meant that I needed to revisit the editor and have a play later to truly understand what I was doing. We built on the first simple SELECT statement, adding to it line by line and running it to make sure the results (all Jane Austen related) were good before building it up further, querying first using strings then moving onto URIs. We had to be very specific if we wanted to return results. Owen explains how this need for specificity demonstrates how SPARQL is not so good for searching but is very good at precise retrieval.

As the next section 'What's wrong with MARC' was introduced I overheard a couple of comments suggesting that there might not be anything wrong with it, but after Tom

calmly and clearly exposed its failures to us one by one, I knew no cataloguer in the room could justify wanting to hold onto it. He started with a slide of a catalogue card pre AACR2 and everyone began to sit up and pay attention. Data is organised in OPACs with labels down the side saying what each bit is. But with a catalogue card it's the order that the data is laid out in that tells us what it means and its place in the drawer. You understand it because of where it is in the record and how it is separated by punctuation e.g. a comma tells us 1965 isn't part of London in 'London, 1965.' But if you take a bit out of context or out of the record it doesn't make any sense because it's not self-describing. So from these card catalogues MARC was born and so it really wasn't designed for use in OPACS. We have to fight against it for example if you want a title to display without a statement of responsibility you're going to have to muck about with it to get it stripped out of the 245 field. It was designed for human comprehension not for machine comprehension, and even cataloguing with RDA in MARC hasn't changed much about the way it is stored.

We were then shown a .mrc record with MARC in its 'natural form' to make the distinction that RDF is not hard and that MARC in its raw state is not easily readable anyway. Everyone was all quiet – I think it worked well as it showed us how complex the MARC encoding is that everyone has got used to working with, and showed that if we can cope with that then we could learn to cope with something like Linked Data. Tom then explained how MARC is for storage, manipulation, display, input, exchange and distribution, publication, and is the lingua franca of library cataloguing he told us that Linked Data shouldn't take on all these roles and none of us would be expected to sit and type out RDF. What followed was an in depth analysis of specific issues. I'll explain some of them briefly…

- Information is duplicated e.g. languages could be in 008, 041, 240, 546 fields.
- What should be unique identifiers for author names are written in various ways with different dates and various punctuation and it is problematic when they change for example when a person dies.
- Text processing is slow when we need to do things like strip out punctuation.
- Information in a 700 field is meaningless without a relator term.

- RDF triples can be isolated and still remain meaningful but with MARC sections of data have no other context.
- Only libraries use MARC so we are tied to library specific software and processes and outside agencies can't take advantage of library data and standards.
- The issues with finite notations and too few indicator fields prompted Tom to jokingly mention SuperMARC because it has 4 character tags, 3 indicators, and character codes of 3 figures. I got the feeling this is not where we want to be headed.

By this point in the day we all had a lot to think about and of course it's not going to be too comfortable to accept that MARC's days are numbered. I had all MARC's failings swimming round in my head when Alan Danskin began his demonstration of 'RIMMF RDA in many metadata formats.' In comparison to MARC this looked fantastic to me but we were soon to learn it's not a tool for cataloguing but a tool for training and visualising RDA. It shows FRBR works, expressions and manifestations and sticking to the Jane Austen theme Alan demonstrated how to create a record for Emma. I liked the way fields were pre-populated with content, media and carrier depending on the template chosen and I appreciated the drop down menus that appeared each time you typed into a field enabling you to select text that has already been entered into the dataset. The system generated composite key builds up as you add more information to the record and acts as a unique identifier for the thing in the database. For each new authorised access point included, a link to the relevant section in the RDA tool kit is added and, unlike MARC, you can generate effective retrieval based on the relationships between works. Clearly a proper understanding of FRBR and RDA is important for new (and experienced) cataloguers but I think I would be more inclined to use RIMMF for this purpose if it allowed me to create records that could actually be imported into my own library management system. Maybe future realisations of the software will have more functionality in this area.

Finally it was time for the bit we'd all been waiting for and Tom's discussion of BIBFRAME. He explained how the BIBliographic FRAMEwork initiative is the Library

of Congress' attempt at demonstrating credible progress towards a replacement of MARC. By this point in the day I was feeling super inspired and positive about Linked Data and was excited by the prospect of learning more and preparing for it to feature in the future of cataloguing. I felt I had understood the main principles including reading RDF and grasping SPARQL. When it came to BIBFRAME, however, I felt like I had gone back to the beginning again. It might have been because it was the last presentation of the day but I found it difficult to comprehend and I couldn't easily connect the BIBFRAME model with the things I had learnt so far. It left me wondering whether 'replacing MARC' with another similarly closed, complex, library defined system such as BIBFRAME would truly allow us to make the most of the benefits of Linked Data.

The final comments of the day gave us some tangible ideas to take back to our institutions. For a start how much can we reasonably demand from our LMS vendors? Owen reminded us that Open Source software provided by Koha and Evergreen for example will often support Linked Data but the large proprietary vendors are not there yet. Whilst they don't yet offer systems that support it, we should expect our LMS providers to be active in the Linked Data community and to contribute to the discussion. Warning bells should ring if the subject is ignored by a vendor completely. We should also understand that when creating tender documents it may not be enough to simply request that a LMS 'supports Linked Data' as the vendor can easily say that it does by mentioning something to do with RDF but in reality this may mean a minimum amount of functionality. Despite this, as Céline Carty pointed out, there is still value in asking the question for unless it is asked then the vendors will not develop anything, as they will assume there is no demand.

Overall I came away from the event feeling pretty good and having learnt a huge amount. It's clear that we, as cataloguers, have a responsibility to educate ourselves about the Linked Data landscape and the worst thing we could do now would be just to wait and let it happen to us in whatever prescribed form 'it' may take. We have to make sure we are stakeholders in the systems of the future. This may mean critically engaging with the BIBFRAME debate or perhaps it may mean supporting

our own Linked Data projects in house, but either way cataloguers should be part of the discussion. For me, I know I have a lot more to learn but I like to be prepared and as we are clearly on a road to Linked Open Data I'm going to make sure I stay fully engaged.

Twitter storify:
https://storify.com/CILIPCIG/cilip-cig-linked-data-event-february-2015

The presentations:
http://www.cilip.org.uk/cataloguing-indexing-group/presentations/linked-data-what-cataloguers-need-know-2015