Yes, yes, of course. Linked Data. It’s… you know… data that is linked…I feel a bit of a fraud writing for an issue of C&I that is all about linked data. I am a linked data beginner. I could nod sagely when reading that the BBC used linked data to help with their Olympics coverage in 2012 but, if asked to explain what it meant, I was pretty much stuck with vague hand-waving explanations. When I helped Tom and Owen organise the CIG Linked Data event in November, my motivation was that I wanted to learn more myself.

In my years as a cataloguer, I have grown accustomed to non-cataloguing colleagues joking with me about how mysterious the workings of cataloguing seem to them, full of jargon and acronyms that sound like an opaque foreign language (MARC, Z39.50, AACR2, RDA, ISBD). I have to say that I’ve sometimes felt that same feeling of being in a foreign land – “I don’t understand this language” – when people start talking about the semantic web and linked open data. For anyone who feels the same way, I hope the articles in this issue of C&I help to start make sense of some of this unfamiliar world.

Linked data is about data, the structuring of data, the meaning of data, the uses of data to provide users with information, knowledge and – crucially – connections. These are very much the concerns of the cataloguer. Tom Meehan has written articles based on his extremely informative talks at the Linked Data event explaining what Linked Data is, why it matters to cataloguers.

In the last couple of years, the Library of Congress Bibliographic Framework Initiative (Bibframe) has been looking at ways to replace MARC in the representation and exchange of bibliographic data and has concluded that the answer lies in a linked data approach. Beyond the buzzword, then, there is a value in understanding the principles and potential of linked data, as well as the limitations, in order to follow the discussions about the future of cataloguing. Tom looks at what the problems are with MARC that mean it needs to be replaced at all, as well as bravely attempting to summarise what Bibframe is and where it might lead.

Linked Data isn’t only about future potential. There are many projects in libraries, museums and other cultural heritage institutions using linked data in fascinating ways, connecting people to information and connecting pieces of information to each other in new ways. Peter McKeague describes the SENESCHAL (Semantic ENrichment Enabling Sustainability of arCHAeological Links) project and Corine Deliot talks about the British Library release of the British National Bibliography as linked data. Gopal Dutta also explores a presentation given by Richard Wallis of OCLC about how linked data might be used for ebook discovery. There is much to learn here, for the novice like myself, or those more familiar with Linked Data.

C&I editors
Helen Garner (h.j.garner@shu.ac.uk) or Heather Jardine (heatherjardine402@hotmail.com).
What is Linked Data?

Linked data is a powerful means of structuring and publishing data on the web. The world wide web we are used to is essentially a linked web of documents; linked data provides a means to create a web of data. Rather than a series of locked down catalogue databases using standards specific to libraries, linked data offers the chance to openly share, link, and enrich this data not only with other libraries but beyond libraries and beyond the purposes to which we currently put it. This article will concentrate on what linked data actually looks like and how to read it. This is important to be able to understand conversations about linked data, which is becoming increasingly prominent in libraries and, in initiatives like Bibframe, is seeking to replace MARC as a standard for encoding and exchanging bibliographic data.

Linked

Cataloguers are used to creating bibliographic data. Most of our traditional bibliographic data is recorded in MARC, of which very little is linked at all. There are 856 fields for URLs but these refer mostly to external documents; they do not allow us to find out any more about the data itself. A publisher’s name is recorded in a 260 field subfield $b and there is little more we can find out about that publisher except by taking the text of their name and looking somewhere else for it, e.g. by searching on Google. Where MARC cataloguing does have a strong system of linking is in name and subject authority files. However, these links are established using highly changeable forms of names. These names are frequently internationally established but often also confined to local lists. Other than viewing an authority record, it is also not straightforward for a computer to follow those links to find out more information for the benefit of the user or the cataloguer.

Data

HTML, by contrast, is designed to produce linked and structured documents for humans to read, understand, and follow links from. Below is an excerpt of the HTML source from a WorldCat page for a book, showing a row of a table (tr) with a header cell (th) and a normal cell (td):

```
<tr id="bib-publisher-row">
  <th>Publisher:</th>
  <td id="bib-publisher-cell">Boston, Little, Brown, 1945.</td>
</tr>
```

This table row contains text, not data. A search engine will find this hard to decode, as it has to know that the table header describes a particular book (rather than, say, a recipe for risotto) and refers to the table cell immediately following it. Even then, it may have problems figuring out exactly what the significance of the English text “Publisher:” is, or what exactly it refers to in the context of the document: are there several books described on the page; is “Boston” and “1945” part of the publisher’s name? Another catalogue may instead use “Publisher” or “Publication details” or “Imprint” or a term in French or Chinese.

Much of the web’s power and reach depends on it being linked and largely open but we have relied on clever search engines with mysterious algorithms to make sense of those basically textual documents and extract their meaning. Linked data aims in part to solve this by allowing metadata producers to share and publish data as structured data rather than documents, or alongside documents.

---

Open

The web relies on openness to an extent which most of us take for granted, but a closed web would be largely unworkable. Although it is possible to create and use large amounts of linked data internally, such as to power a web site or an inventory system, there is a strong assumption of openness with linked data. Indeed, links need to be open to be shared and useful. There is also a strong movement for openness, starting with the open software movement and increasingly in open access within academia, and open data is one aspect of this.

The Open Data Institute defines open data as:

“information that is available for anyone to use, for any purpose, at no cost.

Open data has to have a licence that says it is open data. Without a licence, the data can’t be reused.”

The linked data version of the British National Bibliography, for instance, has been made available under a Creative Commons CC0 1.0 Universal Public Domain Dedication licence. This means that the British Library Board makes no copyright, related or neighbouring rights claims to the data and does not apply any restrictions on subsequent use and reuse of the data.

This greatly increases the likelihood this data will be re-used. Even a small qualification to an open licence can severely restrict the use of data or content. A non-commercial licence, for example, raises so many questions about the meaning of “commercial” and re-use that Wikipedia will not accept photographs released as Creative Commons Attribution-NonCommercial (CC BY-NC).

Linked Data standards

Linked data is not a technical standard in itself but embraces a number of principles as set out by Tim Berners-Lee in 2006:

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- Include links to other URIs so that they can discover more things.

HTTP URIs cause much confusion but are in effect just URLs used as identifiers. For instance, http://en.wikipedia.org/wiki/Evelyn_Waugh is a URL for a web page in English about Evelyn Waugh. It tells a browser where to locate the page and is not an identifier. A HTTP URI is much the same in form but different in intent: http://id.loc.gov/authorities/names/n79049248 is a URI coined by the Library of Congress for Evelyn Waugh. If you put the URI in a browser you will be directed to a human-readable equivalent in HTML with a different URL; a computer making the same request will receive a copy of the data in RDF at another URL. What is important is that the URI is an identifier, and preferably one where more information can be found, potentially with further links to yet more resources.

---


RDF is the backbone of linked data and will be the focus of the rest of this article. It stands for Resource Description Framework and is a specification of the World Wide Web Consortium (W3C). The W3C of course also look after web standards like HTML, XML, and CSS. It is important to note that RDF is by no means a library-specific standard. RDF is not a file format or a specific way of marking up data but is rather a model based on representing data in triples. This is best explained by working up an example.

Humans can exchange information with each other using languages like English. For example, we can make an assertion about a book by writing or speaking a simple sentence:

Brideshead Revisited was written by Evelyn Waugh.

We can start by dividing this into Entities and Relationships, similar to the modelling behind FRBR. There is a subject entity (“Brideshead Revisited”), an object entity (“Evelyn Waugh”), and a relationship between the two (“was written by”), normally referred to as the predicate in RDF.

Brideshead Revisited was written by Evelyn Waugh

This is still English text: the entities are textual, unidentified, and not linked. To identify them unambiguously, we can assign them URIs. As we discovered above, the Library of Congress have created URI identifiers for their name authorities: the URI for the work Brideshead Revisited is http://id.loc.gov/authorities/names/no97080492. If we replace the English title with the URI, we get:

http://id.loc.gov/authorities/names/no97080492 was written by Evelyn Waugh

Of course, the Library of Congress have also created a URI for Evelyn Waugh, so we can include that too:

http://id.loc.gov/authorities/names/no97080492 was written by
http://id.loc.gov/authorities/names/n79049248

Library of Congress does not maintain a URI for the “was written by” relationship (setting aside Bibframe for now), but Dublin Core has the creator term, so we can use that. We now have three URIs:

http://id.loc.gov/authorities/names/no97080492
http://purl.org/dc/terms/creator
http://id.loc.gov/authorities/names/n79049248

If we enclose these URIs in angle brackets and add a full stop on the end...

<http://id.loc.gov/authorities/names/no97080492>
<http://purl.org/dc/terms/creator>
<http://id.loc.gov/authorities/names/n79049248>.

...we have a valid piece of RDF! Because there are three pieces to it, it is called a triple. Triples are the basis for all RDF: everything in RDF is expressed in triples. The three pieces are called, somewhat similarly to the sentence we started off with, the Subject, Predicate, and Object. A triple is an assertion which stands on its own. It is not part of a record and, unlike a MARC field, does not need one to make sense, although in practice the principle of putting useful information about something at the URI means that data becomes aggregated into documents analogous to a record. Although triples by themselves are inherently simple, in combination they can express very complex ideas.

---

Turtle

To make RDF easier to read, both for humans and computers, it can be written out in several different ways. The example above is written in a format called *N Triples*, of which more later. The most common format for people to read, and the one this article will prefer, is called **Turtle** [or TTL, for Terse Triple Language]. Our example in Turtle would look like this:

```turtle
@prefix lc_names: <http://id.loc.gov/authorities/names/> .
@prefix dc: <http://purl.org/dc/terms/> .
lc_names:no97080492 dc:creator lc_names:n79049248 .
```

The first two lines are actually an attempt to make this easier to read! After the word @prefix, a prefix is supplied. This could be anything: it's just to make it easier to read. Lastly there is the base of a URI, followed by the full-stop. Whenever you see that prefix in the document, you can substitute it with the base of the URI. There is still only one triple in the example but now it fits all on one line. The benefits of Turtle become more obvious when we make more assertions about Brideshead Revisited by adding more triples:

```turtle
@prefix lc_names: <http://id.loc.gov/authorities/names/> .
@prefix lc_languages: <http://id.loc.gov/vocabulary/languages> .
@prefix dc: <http://purl.org/dc/terms/> .
lc_names:no97080492 dc:creator lc_names:n79049248 ;
   dc:created "1945" ;
   dc:extent "1 volume" ;
   dc:language lc_languages:eng ;
   dc:title "Brideshead revisited" ;
```

Because all six triples have the same subject— the URI for the book— it is not repeated. More Dublin Core properties are used: *created* for the date of creation or publication, *extent, language, title* and *type*. However, for the language we have used the linked data version of the MARC language codes to provide us with a URI. For the type we have used a separate vocabulary maintained by Dublin Core; as it has only been used once in this example, the full URI has been used, although there is nothing stopping us using a @prefix statement for that too. Several data elements here have been expressed as literals. In other words, the values themselves are recorded directly in quotes. Literal values are practically universal in MARC where even the headings are records as literal strings rather than links or ID numbers. As its very name suggests, linked data generally tries to avoid them in preference to URIs identifying and linking to data. However, even a system based on URIs needs to have some literal data in it at some to make the data meaningful.
Much like a HTML document, though, what makes linked data powerful is the ability to follow links. If we put URI for Evelyn Waugh- http://id.loc.gov/authorities/names/n79049248 – into a browser, we get a page of HTML:

This is because the server knew that we were using a web browser so gave us a page we could read. The URI we typed in has changed to a URL for this page: note that it ends in "html". A computer programme might make a similar request for the URI and get some raw RDF back. Indeed, further down the page is a list of other formats, one of which is N Triples, which we encountered above. Below is an excerpt:
Note first of all the slightly different URL. Each line of the document is a separate triple and, as such, each triple is relatively easy to read: the first two lines assert what type of thing http://id.loc.gov/authorities/names/n79049248 is, the third gives an authorised form of his name. However, the very long lines and lack of any helpful formatting make it hard to see what is going on; below is a small excerpt of this data converted to Turtle:

```
@prefix lc_names: <http://id.loc.gov/authorities/names/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix mads: <http://www.loc.gov/mads/rdf/v1#> .
@prefix viaf: <http://viaf.org/viaf/sourceID/> .

lc_names:n79049248  rdf:type mads:PersonalName ;
              rdf:type mads:Authority ;
              mads:authoritativeLabel "Waugh, Evelyn, 1903-1966"@en ;
              mads:hasExactExternalAuthority viaf:68937142 .
```

Now at least we can see that those types- using the MADS schema- are personal name and authority; to find out more about what those mean, we can follow those URIs. The authoritative label in the third triple is the LC authorised form; the "@en" after it signifies that the literal value is in English. This convention allows for multiple values in different languages. The fourth triple asserts that Evelyn Waugh has an external authority identified by the URI http://viaf.org/viaf/sourceID/68937142. We, or a computer programme, can continue to follow such URIs and find more and more information out. For instance, the Virtual International Authority File has links to a variety of national authority schemes as well as Wikipedia. This holds out the possibility of enriching catalogue
displays with biographical information or, going the other way, embedding bibliographic data or catalogue searches into Wikipedia.

Serializations

As discussed above, RDF can be expressed in several different ways. These are called serializations. They are all readable by computers but have differing advantages in various contexts. The most human-readable one is **Turtle**, which can be seen in the last example, repeated below:

```
@prefix lc_names: <http://id.loc.gov/authorities/names/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix mads: <http://www.loc.gov/mads/rdf/v1#>.
@prefix viaf: <http://viaf.org/viaf/sourceID/>.

lc_names:n79049248 rdfs:type mads:PersonalName ;
    rdfs:comment mads:authoritativeLabel "Waugh, Evelyn, 1903-1966"@en ;
    mads:hasExactExternalAuthority viaf:68937142 .
```

It is also relatively easy to write out Turtle by hand once you have established some prefixes, although the syntax can become complicated. **Notation3**, or N3, is closely related and looks very similar.

The first serialization used in this article was **N Triples**. The last example re-written as N Triples looks like this:

```
```

There is no abbreviation, no prefixes, and no attempt to make it easier to read. It is far easier, however, to see the structure of the underlying triples on each line even if it is very difficult to fit each triple on a separate line!

**RDF/XML** is the most commonly seen form of RDF. Indeed, there is a common misconception that RDF has to be in XML. This is not the case although RDF/XML was the first serialization approved for use with RDF. Here are the same four triples written as RDF/XML:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:lc_names="http://id.loc.gov/authorities/names/"
    xmlns:mads="http://www.loc.gov/mads/rdf/v1#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:viaf="http://viaf.org/viaf/sourceID/"
>
    <mads:PersonalName rdf:about="http://id.loc.gov/authorities/names/n79049248">
        <mads:hasExactExternalAuthority rdf:resource="http://viaf.org/viaf/sourceID/68937142"/>
    </mads:PersonalName>
</rdf:RDF>
```

Clearly this is not straightforward to read and the triple structure is largely obscured. However, the ability to process XML is well developed in programming languages and applications so this can be a convenient way of publishing and consuming linked data.
RDF/JSON uses a syntax similar to that used by the programming language Javascript (JSON stands for JavaScript Object Notation) and similar programming languages. This makes it appealing to programmers working in those languages.

```json
{
  "http://id.loc.gov/authorities/names/n79049248": {
    "http://www.loc.gov/mads/rdf/v1#hasExactExternalAuthority": [
    {
      "type": "uri",
      "value": "http://viaf.org/viaf/sourceID/68937142"
    }
  ],
  "http://www.w3.org/1999/02/22-rdf-syntax-ns#type": [
    {
      "type": "uri",
      "value": "http://www.loc.gov/mads/rdf/v1#Authority"
    },
    {
      "type": "uri",
      "value": "http://www.loc.gov/mads/rdf/v1#PersonalName"
    }
  ],
  "http://www.loc.gov/mads/rdf/v1#authoritativeLabel": [
    {
      "lang": "en",
      "type": "literal",
      "value": "Waugh, Evelyn, 1903-1966"
    }
  ]
}
```

There is a further serialization gaining currency called JSON-LD (JSON for Linked Data) which looks similar but is quite different.

**RDFa, Microdata, and Schema.org**

We started the article with an example of bibliographic data expressed unsatisfactorily as HTML. **RDFa** provides a way of embedding RDF data directly in web pages. Our four-triple example expressed using RDFa in HTML `div` elements might look as follows:

```html
<div xmlns="http://www.w3.org/1999/xhtml"
     prefix=""
     rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
     mads: http://www.loc.gov/mads/rdf/v1#
     rdfs: http://www.w3.org/2000/01/rdf-schema#" >
  <div typeof="mads:PersonalName" about="http://id.loc.gov/authorities/names/n79049248">
    <div rel="rdf:type" resource="http://www.loc.gov/mads/rdf/v1#Authority"></div>
    <div property="mads:authoritativeLabel" xml:lang="en" content="Waugh, Evelyn, 1903-1966"></div>
  </div>
  <div rel="mads:hasExactExternalAuthority" resource="http://viaf.org/viaf/sourceID/68937142"></div>
</div>
```
A computer programme reading this web page would be able to extract the underlying triples and therefore the meaning from this. This is highly significant as that computer programme could be a search engine like Google. Instead of having to guess what a web page or an element is about, a search engine would have access to metadata in RDF. Indeed, the Worldcat page from which we took the HTML at the beginning of this article also has linked data embedded in it. At the bottom of the page is a Linked Data section: click on the plus sign (+) and a basic HTML view of it is shown:

This uses the Schema.org vocabulary, with some extensions, in RDFa. Schema.org was set up by the search companies Bing, Google, Yahoo!, and Yandex as a standard vocabulary to be used in web pages, in no way limited to libraries. Clearly there is therefore some benefit in including structured linked data into catalogue pages if they are more easily found by search engines, and therefore researchers. Below is a snippet of what a small number of the underlying triples look like in N3:

---

8. Extracted using Alex Stoltz’s RDF Translator at http://rdf-translator.appspot.com. This was also used for many of the translations and extractions of RDF in this article.
The first triple says this is a book in the schema.org vocabulary: using “a” on its own as a predicate is a very common shorthand for the RDFS “type” property, i.e. this particular subject is a “type” book. The second triple uses a library extension to schema.org to specify the OCLC number. The third triple is more complicated and runs across several lines. The square brackets allow several statements to be made about the place of publication without having to give it a URI of its own. Another way of writing this out with the same meaning and more explicit triples would be as follows:

Here we have given the place of publication an arbitrary name called a blank node. We can then make additional statements about this place: the first says that _:bnode001 is indeed a place; the second that it’s name is “Boston”. Blank nodes are extremely useful for expressing complex concepts in otherwise very simple triples.

**Microdata** is a similar approach to RDFa for embedding metadata into web pages.

**Conclusion: More Than One Way To Do It**

Linked data using RDF is a framework for sharing data over the web, not a library-specific vocabulary. It has enormous potential for sharing library data with the world as well as making library data part of a wider web of data. It is powerful, provides precision, but also needs some definition and agreement to work globally. The last example below demonstrates this:

The same assertion is here made seven times using four different URIs to express the creation relationship

- dc:creator
- schema:creator
- rdau:P60447 (the RDA URI for the creator relationship)
- example:author (one I made up)
and URIs for Evelyn Waugh from five different sources:

- Cambridge
- BNB
- VIAF
- LC
- example (one I made up)

as well as text via a blank node. To be usable the library community will have to choose which ones to re-use and share with others engaged in similar work or whether to make one up a whole new scheme itself. There are clear examples, such as the BNB\(^9\) and Bibframe\(^{10}\), of both approaches. There are great possibilities ahead.

**References**


W3C. *Resource Description Framework (RDF)*. http://www.w3.org/RDF/

---


Introduction

This paper describes the development of a linked data instance of the British National Bibliography (BNB) by the British Library\(^1\). The focus is on the development of an RDF (Resource Description Framework) data model and the technical process to convert MARC 21 Bibliographic Data to Linked Data using existing resources. BNB was launched as linked open data in 2011 on a Talis platform. In 2013 it was migrated to a new platform, hosted by TSO. The paper discusses issues arising from the development, implementation and running of a linked data service. It also looks ahead to plans for future developments.

Motivations and approach

The British Library released the British National Bibliography (BNB) as Linked Open Data in July 2011. Our motivations were twofold. Firstly, there has been an increasing commitment from the UK Government since 2009 to the principle of opening up public data for wider re-use. The Linked Open BNB forms part of the Library’s response to this agenda. One of our aims was to break away from library specific formats and use more cross domain XML-based standards in order to reach audiences beyond the library world. Secondly, we wanted to be part of the Linked Data conversation. We wanted to experiment and see what it meant to publish bibliographic data as Linked Data; there were many claims made about the benefits of Linked Data and we were hoping to see them tested.

Our approach was pragmatic. We didn’t try to model the whole bibliographic universe, but chose a particular data set, the British National Bibliography. We kept in mind other datasets in order to ensure that the decisions we made would be extensible but we primarily modelled the BNB.

There were many reasons why we chose to offer the BNB for this experiment. Firstly, this dataset is an authoritative source of information about UK publications from 1950 to the present; it is a general database of published output and not an institutional catalogue of unique items. This made it suitable for re-use. Secondly, the data is reasonably consistent and well maintained. The records all have Dewey Decimal Classification (DDC) numbers and headings are generally under authority control. There are of course some caveats to add. As anybody who has worked with MARC data will realize, the data is not as consistent as we would wish. BNB data was not created for machine actionability; changes in policy and cataloguing standards as well human error over the lifetime of the dataset means that our options are sometimes constrained by the data. Another reason we chose the BNB is that it represents a significant amount of data — about 3 million records in several languages. Lastly, the rights attached to this dataset were clear. Where we have not created the metadata ourselves, the Library has secured the rights to distribute it in perpetuity. We were therefore able to make the BNB available under the Creative Commons licence CC0. During its lifetime, the BNB has migrated from one platform to another as technology has developed. It began in print, which was supplemented by magnetic tape and then CD-ROM. It was made available via Z39.50 and on the Web; linked data is just the next technology.

Our intention was to discover how much could be done using our existing resources in staff, systems and knowledge. Metadata Services staff involved in the project have developed expertise in bibliographic standards and tools to manage and manipulate large volumes of bibliographic data, predominantly in MARC and also had some experience of HTML, XML and XSLT but the team did not include programmers or data architects. Training in RDF and the principles of data modelling was provided by Talis, who were also our consultants and mentors throughout the project. Some new tools were developed, for example to generate Uniform Resource Identifiers (URIs) or to be able to link to external linked datasets, but on the whole we were able to use our existing tools.

---

\(^1\)This is a companion to the presentation available on the CIG website at: [http://www.cilip.org.uk/cataloguing-and-indexing-group/linked-data-what-cataloguers-need-know](http://www.cilip.org.uk/cataloguing-and-indexing-group/linked-data-what-cataloguers-need-know)
The data model and the modelling process

Modelling first involved identifying our objects of interest, which means stepping back from MARC to identify what the catalogue record says about “things in the world”. These include concepts and abstractions as well as material objects, for example bibliographic resources, persons, organizations, places, subjects, etc.

In order to identify these entities we had to assign URIs. This is more complex than it sounds and involves a number of decisions. We chose to mint our own URIs for most of our entities rather than rely on external sources. For example, we created our own identifier for William Shakespeare rather than rely on the VIAF ID. There are two reasons for taking this approach. Firstly, however authoritative the data source, there is no guarantee that it will always be available. Secondly, the external linked data set may not include all of the resources we want to make statements about. We also discussed whether we should opt for opaque or human-readable (“transparent”) URIs. Transparent URIs are easier to work with because the ID reflects the underlying semantics, but there is an argument that, in a multilingual environment, opaque URIs are more inclusive. We discussed the patterns that the URIs should follow and applied, as far as possible, guidance provided by the Chief Technology Officer Council in its report “Designing URI Sets for the UK Public Sector”. Finally, we had to consider how to produce valid URIs, i.e. conformant with the URI syntax specified by IETF.

The next step in the process involved describing those entities and how they relate to each other. Our approach was to use classes and properties from existing RDF vocabularies as much as possible. We looked to see which ontologies other LOD projects were using at the time and settled on a mix of Dublin Core, The Bibliographic Ontology, FOAF: Friend of a Friend, the Event Ontology, etc. We tried to use library-domain ontologies sparingly because, as previously mentioned, we were trying to reach audiences beyond the library world. This is also one of the reasons why in this first instantiation of the Linked Open BNB we did not use the FRBR (Functional Requirements for Bibliographic Records) ontology. There were also two other reasons for this: firstly, we would have had to do a lot of work upfront to identify the FRBR entities in our MARC records and we simply had not got the time as we were working to a tight deadline. Secondly, there were some differences of opinion between ourselves and Talis developers, who viewed the FRBR model as too complex. However, we did not eschew library-domain ontologies completely. We found the ISBD element set particularly useful as many of the properties we needed were defined with few constraints, especially with respect to expected values (range). There was no class specified as the expected value, which made them ideal to record a number of free-text MARC notes. They also provided granularity, thus enabling us to avoid mapping all of these to a generic dcterms:description.

In some cases, we have chosen to use classes that appear to duplicate each other. For example, org:Organization is defined as owl:equivalentClass to foaf:Organization similarly for foaf:Agent and dcterms:Agent. By using classes from different schemas our dataset can mesh with a broader range of datasets, i.e. those which only use org:Organization and those which only use foaf:Organization. Linked data applications of varying degrees of sophistication can consume our data more easily as they do not necessarily need the capability to support particular reasoning and inference rules.

4 http://tools.ietf.org/html/rfc3986
5 For a full list of British Library URI patterns, see http://www.bl.uk/bibliographic/pdfs/british_library_uri_patterns.pdf
6 A full list of the ontologies used is available from: http://www.bl.uk/bibliographic/datafree.html
7 International Standard Bibliographic Description http://metadataregistry.org/schema/show/id/25.html
8 http://www.w3.org/TR/vocab-org/
9 http://xmlns.com/foaf/spec/
10 A similar argument can be made about properties. For further details on this topic, see Pete Johnston’s blog entry http://archiveshub.ac.uk/locah/tag/vocabulary/, in particular the section on Inferencing.
We also defined our own classes and properties, documented in the British Library Terms RDF schema,\(^\text{11}\) where necessary. There were two circumstances in which we decided it would be appropriate to define our own terms. Firstly, if we were unable to find a property of sufficient granularity to record a piece of data we needed. An example of this is blt:bnb to record the BNB number, which we preferred to the less specific dcterms:identifier. The other circumstance was if the class/property was required by a specific feature of the model. An example of this is our modelling of the publication statement as an event.

We also created some classes and properties in order to facilitate searching: for example, we created the classes blt:TopicLCSH and blt:TopicDDC as sub-classes of skos:Concept. These enable users to request a more refined search based on a particular LCSH subject or DDC number. We also created inverse properties to facilitate navigating from one resource to the other: for example blt:hasCreated as inverse property of dcterms:creator as well as blt:hasContributedTo as an inverse of dcterms:contributor. This makes it easier to query the data and facilitates the retrieval of all resources created or contributed to by a particular entity. Overall, we created relatively few classes and properties; our priority was to re-use existing ontologies. Re-using metadata facilitates interoperability and minimizes the burden of maintaining our own metadata.

The outcome of this modelling activity is illustrated by two diagrams, one for books\(^\text{12}\) and the other for serials\(^\text{13}\). The models for books and for serials are not fundamentally different, they include different classes and properties as required by the different types of material.

\(^{11}\) [http://www.bl.uk/schemas](http://www.bl.uk/schemas)

\(^{12}\) [http://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf](http://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf)

\(^{13}\) [http://www.bl.uk/bibliographic/pdfs/bldatamodelserial.pdf](http://www.bl.uk/bibliographic/pdfs/bldatamodelserial.pdf)
One of the features of the model is that we decided to model the publication statement as an event. This was motivated by the (future) need to represent forthcoming publications. The BNB includes CIP (Cataloguing in Publication) records, which are advanced notifications of new publications received up to 16 weeks prior to publication. We wanted the event model to be extensible, to model other events in the life of the resource, for example when the book is acquired, launched or goes out of print. Modelling the date, place and publisher as entities meant that they could be identified by URIs, rather than by literals (text strings), which offers the prospect of enhanced retrieval and aggregation of data.

We also made extensive use of the “foaf:focus” property to relate “things in the world” such as people, organizations, places, etc. to their SKOS concepts. This property enables “crossover from the bibliographic and cataloguing facts associated with a particular thesaurus’s conceptualization of an entity to facts and assertions about the entity itself”. This allows the model to make a clear distinction between the real world object and its bibliographic surrogates. For example, “London” the current capital of England and the UK as a single “thing in the world” may be the focus of multiple concepts belonging to different concept schemes, such as LCSH, RAMEAU, etc. This approach is not without its challenges and is still contentious.

From MARC 21 to RDF serializations

We select records from the full BNB file, as we process books and serials separately. The next step is to do a character-set conversion. We hold our MARC records in “decomposed” Unicode/UTF-8 but the RDF recommendation is for literals to be in Unicode Normal Form C, i.e. “composed”. The data is then normalized for improved matching and transformation using MARC Global, a tool developed by TMQ (The MARC of Quality). British Library-coined URIs are generated in the MARC records and URIs linking to external datasets such as VIAF and LCSH are added. All this processing is done with a suite of British Library tools called Catalogue Bridge utilities. The enhanced MARC file is then converted to RDF/XML using XSLT. After quality checking the RDF/XML with an open source tool, Jena Eyeball, the resulting Linked Open BNB file is converted to N-Triples for loading onto our hosting platform. Data dumps in both serializations (RDF/XML and N-Triples) are also loaded to our Downloads page.

With respect to linking to external datasets, we chose general resources to give our data broader context, such as Geonames (for country of publication), Lexvo for languages. We also linked to library domain datasets: VIAF and LCSH, as already noted, but also Dewey.info and the MARC Country and Language codes. One technique involved generating URIs automatically from record data. For example, to link to Dewey.info, which provides access to the top three levels of DDC, all digits after the decimal point in the Dewey number 641.5686 were stripped and the remaining

14 http://wiki.foaf-project.org/w/term_focus (written prior to the creation of this property)
15 For further details, see Peter Johnston’s blog post at: http://efoundations.typepad.com/efoundations/2011/09/things-their-conceptualisations-skos-foaffocus-modelling-choices.html
16 http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#section-Literals section 3.4
17 http://www.marcofquality.com/
18 http://viaf.org/
19 http://id.loc.gov/authorities/subjects.html
20 http://www.w3.org/TR/xslt
21 http://jena.sourceforge.net/Eyeball/
22 http://bnb.data.bl.uk
23 http://www.bl.uk/bibliographic/download.html
24 http://www.geonames.org/ontology/documentation.html
25 http://www.lexvo.org/
26 http://dewey.info/
27 http://id.loc.gov/vocabulary/countries.html
28 http://id.loc.gov/vocabulary/languages.html
Dewey number inserted in a URI of the form http://dewey.info/class/641/. Other techniques included automatic matching of authorized headings in our bibliographic records with the appropriate strings in linked data dumps and retrieving the corresponding URIs; this was used for LCSH and VIAF. Finally we also used cross walk matching for coded data.

Current outcomes and future developments

Work began on the project in late 2010 and the Linked Open BNB was initially launched on a Talis-hosted platform in summer 2011. The Talis platform offered a range of options for querying the data, including a SPARQL endpoint. In the first year of operation, the number of hits against our SPARQL endpoint increased from 38,000 in the first month to over 9,000,000. Discussing how to analyse and derive value from these usage statistics (e.g. sources, types of queries) came to an end in July 2012 when Talis announced its withdrawal from the Semantic Web business, because the market was developing more slowly that they had anticipated.

The British Library selected TSO to host the Linked Open BNB when the contract with Talis expired. Data and services were migrated over a couple of months and went live in July 2013. Two BNB datasets, Books and Serials, (both with VoID descriptions) are now offered by the Library at http://bnb.data.bl.uk. The data can be queried from the SPARQL endpoint (http://bnb.data.bl.uk/sparql) and a SPARQL editor (http://bnb.data.bl.uk/flint) is also available. The BNB is refreshed each month, both on the platform and on the British Library bulk downloads page.

Metadata Services tracks usage of the BNB by monitoring the number of hits on the SPARQL endpoint and recording the number of bulk downloads from the British Library website. We have supplied the Linked Open BNB for use by internal and external projects, e.g. as test data for a British Library semantic search demonstrator and to assist Microsoft in their research into linking structured data. We also have anecdotal evidence of usage – such as references to the dataset by third parties. However, it is undeniable that detailed assessment of third party usage is a problem area for all organizations involved in offering linked open data services. We are working with TSO and the UK Government Open Data Forum to develop better metrics and impact assessment techniques.

We are also looking to develop our linked open data offering and have identified a number of areas for investigation. Firstly, refining and extending the model. Ideas so far include modelling entities currently excluded such as conferences or forthcoming publications. Secondly, enriching the Linked Open BNB with links to other resources. Potential candidates for linking include the International Standard Name Identifier (ISNI), LC/NACO and DBpedia. We have explored the feasibility of making more granular links to places, by matching place of publication data with Geonames at city level. Unfortunately, this is one of those situations where MARC data is not normalized or clearly disambiguated, so the process requires more manual quality assurance than we can currently afford.

More fruitful in the short term may be linking to other national bibliographies and we are in the process of linking BNB and bibliographic resources in the Deutsche Nationalbibliothek.

In the longer term, we intend to revisit FRBRization of the data. When we first started the project, FRBRizing was deemed out of scope as we were aiming to reach a new audience, one beyond the library community. However, it has become clear that there are different communities of users out there with a variety of needs and that two versions of the Linked Open BNB, one with the current model and one with a FRBR model would not be mutually exclusive.

---

29 http://www.information-age.com/technology/information-management/2111803/talis-shuts-down-semantic-web-operations%C2%A0
30 http://www.tso.co.uk/our-expertise/technology/openup-platform
31 VoID is an RDF schema vocabulary for expressing metadata about RDF datasets. See http://www.w3.org/TR/void/ for further details.
32 http://www.bl.uk/bibliographic/download.html
33 The white paper is available here: http://research.microsoft.com/apps/pubs/?id=193076
Encouraging more use of the data is also on our list of priorities. As part of that, we are gathering feedback on how to improve the information provided on our documentation pages\(^\text{34}\), especially that provided for developers.

Finally, we are looking to expand the scope beyond the BNB. Sheet music is also covered by UK legal deposit legislation and the BNB was complemented for many years by the *British Catalogue of Music*, so we are currently exploring the feasibility of publishing sheet music as linked open data.

### Challenges and benefits

The many challenges of the project can be summed up by “Converting MARC 21 records is challenging!” It is not surprising that over the 60+ years of BNB’s existence there have been changes of technology, cataloguing policy and standards as a result of which the data is not as consistent as we would like. We also uncovered character set issues in the legacy data. We were also constrained by the technology we used. For example, generating URIs from (potentially volatile) strings, such as name headings, may result in duplication in the future. Other challenges resulted from our own choices. Modelling the publication statement as an event was more complicated that treating it as a literal.

As a newcomer in a developing technology it was sometimes difficult to know what the best way forward was. In 2010-11, publishing a dataset as linked open data was fairly new in the library domain. There was (and, some would argue, there is still) little consensus on many issues. Whether to use opaque or transparent URIs? Re-use existing ontologies or create your own? Are inverse properties necessary? Should we use the foaf:focus property?

Whilst there were many challenges en route, it is clear that there are also benefits in publishing a dataset as linked open data. Perhaps the most obvious one is that Metadata Services staff learnt a lot about the practical aspects of publishing linked data. The project enabled us to get to grips with some of the more abstract aspects of RDF, data modelling, and identification. It was also a period of intense professional development, which definitely spiced up our working lives! The project took us into the new environment of the Semantic Web and we improved our legacy data.

The data model we developed has received considerable attention inside and outside the Library. The Stanford Linked Data Workshop Technology Plan\(^\text{35}\) recognized the value of our approach. Our colleagues at the Danish Bibliographic Centre (DBC)\(^\text{36}\) decided to re-use and extend the model to describe their own national bibliography. It is fair to say that the project raised the profile of the Library externally and the profile of Metadata Services internally.

In conclusion, the combination of the Library’s bibliographic assets and Talis’s linked data expertise overcame the challenge of publishing the BNB as Linked Open Data. The successful migration to the TSO platform illustrates a fundamentally sound approach, on which we now intend to build.

**Acknowledgement:** Many thanks to my colleague Alan Danskin for reviewing this article. Any inaccuracies are my own.

---

\(^{34}\) [http://bnb.data.bl.uk/docs](http://bnb.data.bl.uk/docs)

\(^{35}\) [http://www.clir.org/pubs/reports/pub152/LDWTechDraft_ver1.0final_111230.pdf](http://www.clir.org/pubs/reports/pub152/LDWTechDraft_ver1.0final_111230.pdf)

\(^{36}\) [http://opensource.dbc.dk/linked-data](http://opensource.dbc.dk/linked-data)
This paper expands upon a presentation given at the CIGS workshop on Linked Data, held in Edinburgh on 18th November 2013. It starts by reminding the reader about the vision behind the semantic web and Linked Data, by looking back at the roots of these developments. I then describe some of my own research work on extracting Linked Data from free text documents, as a way of indexing their content and – ideally – summarising their meaning. Finally the paper examines some of the issues and pitfalls around Linked Data projects, looking at how we can actually make connections between our data and remote but related data. I suggest that the basic principle is analogous to using authority files, but one of the problems is that when “Anyone can say anything about anything” it may be hard to tell which of the many authorities is the one to trust.

I should probably mention at this point that, although I used to work in a National Library, I am a geek not a librarian, so I ask the reader to forgive any blunders made in talking of library matters.

The Semantic Web Vision

Things can move very fast in computer technology. The web – that is to say the “document web”, which we can here contrast with the Data Web (the name Tim Berners-Lee says he should probably have coined instead of “semantic web”) – grew explosively in its first few years. It's hard to believe that before the 1990s there was no World Wide Web. Of course the Internet had been in use for decades, but it was HTML that turned it into a vehicle for the information revolution we are experiencing. By the end of the 1990s we had all realised that creating a web presence for our organisations was not “nice to have” but essential.

In comparison, the semantic web has been a very slow burner. The first version of RDF became a W3C Recommendation in the 1990s and Tim Berners-Lee started talking publicly about the semantic web in 1999. It was back in 2001 that the now famous Scientific American article appeared (Berners-Lee et al. 2001) in which the authors proposed a web of data intended to be read by software agents, not humans. Alongside the raw data (in RDF) the agents would use structured vocabularies and inference rules that would enable them to examine large quantities of information and draw conclusions that could be justified to the human user if required. As Sir Tim put it: “The day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machine, leaving humans to provide the inspiration and intuition”.  

Six years after the first Scientific American article a follow-up appeared (Feigenbaum et al., 2007) which argued that truly significant progress had been made, though the goals described seemed less ambitious than those in the original vision. Another seven years have passed and the semantic web is still hardly a household term. The somewhat ambiguous “Linked Data” label has become fashionable, and one can now say, cautiously, that the movement towards a web of data is at last gaining momentum. In the library and archive worlds, more and more bodies are feeling that it is time to start experimenting with publishing their holdings as RDF triples, with “Cool URIs” that will hook them into the growing web of Linked Data.

---

1 Tim Berners-Lee's famous remark about how the semantic web is supposed to work (incorporated in W3C RDF Concepts, 2002 draft).
3 Whether Linked Data has to involve RDF or not is a question guaranteed to produce lengthy arguments.
Bringing Free Text Content into the Data Web

There are plenty of exciting projects going on that aim to convert structured databases to RDF for Linked Data purposes. I want to consider unstructured information, viz text documents. After all, whilst indexes and field lists are obviously vital for finding information, we know that the really interesting facts are often contained in free text notes or documents that are associated with the database records. If machines are going to be able to make inferences about the knowledge in our datasets they will have to be able to “read” text in some way analogous to how humans do it. My research has involved trying to extract the factual statements embodied in texts (taken from an archaeological site archive), and turn them into RDF triples that could easily be added to the RDF graph derived from a structured database.

To illustrate the process, let us consider the following text:

“In a hole in the ground there lived a hobbit. Not a nasty, dirty, wet hole, filled with the ends of worms and an oozy smell, nor yet a dry, bare, sandy hole with nothing in it to sit down on or to eat: it was a hobbit-hole, and that means comfort.”

RDF triples, consisting of “Subject – Predicate – Object”, can be thought of as simple declarative sentences, where the nodes (the subject and object of a triple) are nouns and the predicate edges (or “arcs”) joining them are verbs. Thus one could represent the first sentence above with the following simple RDF graph of two connected triples:

![RDF graph for the first sentence](image)

Figure 1 shows a graph for the whole of the quotation, involving a graph of five triples. The nodes are nouns and the edges are verbs. The representation is just an example and there are plenty of other ways of expressing the content as triples – just as there are other ways of conveying the same meaning in text sentences.

![RDF graph for the whole quotation](image)

4 This is of course the opening of *The Hobbit*, by J R R Tolkien.
In this example the RDF graph was of course produced a human editor (me) reading the text and deciding what were the key facts to be expressed as triples. I am interested in how well this task can be achieved automatically by software. My `txt2rdf` pipeline reads in text documents and passes them through a sequence of language processing steps (hence the term “pipeline”) including tokenisation, part of speech tagging, named entity recognition and relation extraction. The end result is a set of binary relations between entities mentioned in the text; for example in the *Hobbit* extract, two of the entities detected might be “hole” and “the ground” and, if all goes to plan, the software should assert a relationship between them, namely that the first is located within the second.

Once one has the binary relations between the entities – which are essentially the content-bearing nouns in the text – it is relatively straightforward to assign suitable URIs and turn them into RDF triples. I don't propose to explain the details of the pipeline here, but the interested reader can find them in Byrne and Klein, 2009, which is a paper written for a non-technical audience (that is, for archaeologists interested in processing archive data, rather than for language processing specialists).

The `txt2rdf` pipeline uses statistical techniques and was trained on archaeological data. It performs reasonably well (extracting around 60% of the available “facts” correctly) on data from this domain but very poorly on text from other domains. At present this is still typical of all such tools. The key steps are:

1. Identify and classify the entities mentioned in the text, eg decide whether a given term is a date, a place, a person's name, an archaeological site description, or completely irrelevant. This step is known as Named Entity Recognition or NER.

2. Having found a collection of entities, decide whether there are relationships between them, and if so how these relations should be labelled, eg “site A is located in place B”, “event C occurred on date D”, and so forth. This step is Relation Extraction or RE.

Each of these steps is a standard language processing task and a rough rule of thumb is to expect around 80% success, where “success” is usually measured as the harmonic mean of recall (how many of the available entities the NER step found, say) and precision (how many of those it identified were actually correct). If a pipeline combines NER and RE it multiplies the errors, so the best overall scores to expect are in the region of 64%, which is roughly what my research project produced. With careful training and tweaking these scores could certainly be improved but that means a lot more human effort in writing software and running tests.

The holy grail is to create automatic tools that perform well without having to be carefully and expertly tailored to a particular domain – we are still some way from that. Several research prototypes and a few commercial products have appeared in the past couple of years, for information extraction tasks of this kind. They are generally trained on “news” text – rich in dates, well-known people, etc – and will work best on this kind of material. Table 1 lists some of the currently available tools.\(^5\)

---

\(^5\) These are not recommendations, just a list of systems I am aware of. I haven't evaluated them.
Table 1. Online language engineering tools.

We’ve looked at how language engineering can help to liberate the contents of text documents and turn the facts they contain into RDF triples. This process could also be used for related tasks, such as populating structured database fields, or checking the existing contents of these fields. The relation “event C occurred on date D” can equally well be expressed as a triple, or by putting event C and date D into fields of a database table.

If current tools are still so inaccurate are they of any use? I would say “Yes”, even at current levels of performance. After all, if you are faced with a data cleaning task over millions of records, having a machine do the “easy” 60-80% for you is a very great deal better than nothing. I would expect that, over the next few years, an increasing number of data managers with text-rich archives will turn to language processing tools to assist in their curation tasks.⁶

<table>
<thead>
<tr>
<th>Name</th>
<th>URL</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Calais</td>
<td><a href="http://www.opencalais.com/">http://www.opencalais.com/</a></td>
<td>One of the first on the web. From one of the top research labs. Language engineering tied in to semantic web. Free, but commercial services also available.</td>
</tr>
<tr>
<td>RelFinder</td>
<td><a href="http://www.visualdataweb.org/reffinder.php">http://www.visualdataweb.org/reffinder.php</a></td>
<td>Also from a top research team. Free and open source. Probably requiring a de-</td>
</tr>
<tr>
<td>AlchemyAPI</td>
<td><a href="http://www.alchemyapi.com/">http://www.alchemyapi.com/</a></td>
<td>One of the first commercial products, with some free services. Has a sophisticated online demo on the website to ena-</td>
</tr>
<tr>
<td>TextWise</td>
<td><a href="http://textwise.com/">http://textwise.com/</a></td>
<td>A commercial product, with a simple online demo (paste in your own text) available on the website.</td>
</tr>
<tr>
<td>OpenUp DES</td>
<td><a href="http://openup.tso.co.uk/des">http://openup.tso.co.uk/des</a></td>
<td>Data Enrichment Service. A nice site put up by TSO, based on the well-known GATE tools from Sheffield University. Has a simple demo online that you can paste text into.</td>
</tr>
</tbody>
</table>

⁶ It’s another topic altogether, but “assisted curation” is an interesting problem – where a human and a software agent collaborate on editing data.
Authorities and ontologies

Assuming that we have somehow produced a graph of RDF triples – whether by extracting them from text or by converting a structured database, or even by hand-coding them – how do we make them part of the Linked Data enterprise? I would argue that the fundamental requirement is to hook our RDF data into established public ontologies, much as library catalogue entries are tied to authority records.

The beauty of RDF is that if two graphs share a node they are automatically connected. For example, if my graph happens to include a resource node for J R R Tolkien and so does yours, then our datasets are linked – no protocols, no special software; we’re linked, and a SPARQL query can cross our datasets. But there are a couple of assumptions in that statement. One of the assumptions is that both our datasets are exposed on the web. The other, more germane to this discussion, is that we both use precisely the same data web node for J R R Tolkien. This is in fact very unlikely, unless we take trouble to make it true.

In general, when local data is converted to RDF the first thing that’s done is to decide the “base URI”. I work in the Language Technology Group at Edinburgh University and we have a domain URL of “www.ltg.ed.ac.uk”. When I mint RDF URIs I will prefix them with something based on that. Therefore my J R R Tolkien node might be “http://rdf.ltg.ed.ac.uk/mygraph#jrtolkien”, say. It will clearly not match yours. Therefore, if I am planning ahead I should replace my “default” URI with a canonical one; “http://live.dbpedia.org/resource/J._R._R._Tolkien” would be a good candidate. Assuming you do the same thing in your data, we have Linked Data.

A moment’s thought will show that this process is probably intractable. Every time a triple is generated (and it usually will be generated, by some automatic process) a search has to be made for the best available canonical URI. A simpler procedure is to mint local URIs but to connect them to published standards whenever possible. If there is more than one authority there’s nothing to stop us making multiple connections. In my example, I might add the following triple to my graph:7

mygraph:jrtolkien owl:sameAs dbpedia:J_R_R_Tolkien

If I am the first to publish Linked Data about Tolkien then at this stage I am not connected to anything except the DBpedia node – which will certainly enrich my data, as it gives me access to all the metadata in DBpedia. But as soon as another data publisher comes along and connects to the same resource, I gain a connection to a new remote dataset. As more and more publishers connect to the same shared authority nodes, the web of Linked Data grows automatically. Clearly if we are connected by multiple nodes, not just this single one, then the richness of the network grows and possible SPARQL queries across it can be more sophisticated.

This is an exciting notion, and it’s already happening all around the data world. An example from the cultural heritage field that I work in is the AHRC-funded Seneschal project,8 that has turned a collection of domain thesauri (for monument and artefact types in particular) into RDF ontologies. Projects are underway to link archive datasets that use this standard terminology to the new canonical URIs with a “heritagedata.org” prefix.9

Personally, I think this is great. But it’s not problem free. One of the issues is that there are just too many competing ontologies out there. This doesn't necessarily mean that the network won't get joined up eventually, but it would obviously be more efficient if a query didn't have to step across multiple ontologies to reach a remote data graph. I doubt if anything can be done: the usual result of setting up a committee of the great and good to

---

7 I am following standard conventions for compact URIs, see http://www.w3.org/TR/curie/.
8 http://www.heritagedata.org/blog/about-heritage-data/seneschal/
9 Note that I am talking here of ontologies in the sense of hierarchical terminology lists. There is a separate issue about embedding standard framework schemas in local graphs – much as I used the OWL “sameAs” predicate in the example above.
resolve the problem of having \( x \) competing standards is that one ends up with \( x+1 \) competing standards. A more insidious problem is that, if we decide not to worry about finding canonical URIs as we go along, because we can always sprinkle in some “owl:sameAs” or “skos:exactMatch” links later, we may become lazy about our local RDF design. One comes across some extremely opaque RDF – a long way from the ideal of clean, simple data that generic software agents will read and “understand”. It may become easy enough to link across to a remote dataset – via one of the established stepping stones like DBpedia or geonames – but how will our poor agent find its way about when it gets there, if the local schema is as complex and messy as the relational database it probably grew out of. I think there's a clear case for RDF schema design as an art in its own right.

The subject of schema design is too large a one to start on at this point. But let me finish with what I think is the most pressing requirement. If one accepts my premise that the nodes of the graph are the content-bearing nouns then there is a good case for these being genuinely local, with connections to established authorities whenever possible. But the relationships between them could be made much more generic: one thinks of the “Who?, What?, Where? When?” mantra, that might result in a predicate set like “has agent”, “has classification”, “has location”, “has date”. Readers may see similarities to the Dublin Core set – which at least started out as a laudably simple set of generic properties. If we can make our RDF use as simple and generic a set of predicates as possible, the vision of software agents that can interpret distributed data conveniently for us might just materialise. If we all use locally minted predicates it won't; we'll end up with the same kind of distributed query problems we see now, just based on a different technology.

References


Background

The Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS) identifies, surveys and analyses the historic and built environment of Scotland and its maritime waters. It preserves, cares for and adds information and items in its national collection. It promotes understanding, education and enjoyment through interpretation of the information it collects and the items it looks after.

Our online portal, Canmore, provides searchable public access to a wealth of information on over 320,000 archaeological sites and historic buildings from prehistoric settlements like Skara Brae, Orkney to contemporary architecture such as the Scottish Parliament, with an integrated archive of over 1,254,000 catalogue records. Behind the scenes, information is managed in a relational database. Controlled vocabularies are key to both the storage of information in the database and its discovery online. In particular, we use thesauri to help classify the monument, object and maritime craft types associated with the site records we curate. We encourage the use of thesauri standards amongst local authority Historic Environment Records (HERs) who maintain databases about the historic environment for their Council business, across Scotland and more widely amongst the profession. We also work with colleagues in England and Wales through the Forum on Information Standards in Heritage (http://fishforum.weebly.com/) to promote data standards for heritage.

Drivers for Linked Data

Although we publish our thesauri online as part of Canmore, they are not particularly visible. The thesaurus architecture, implemented in 2005, limits the potential of the terminology as the terms lack the persistent Uniform Resource Identifiers (URIs) that would allow our resources to act as hubs for the Web of Data. For cultural heritage, the original exponents for Linked Open Data come from the research community. The lack of vocabulary control with unique identifiers was recognised in two Arts and Humanities Research Council (AHRC) funded projects (STAR and STELLAR (2)) undertaken by the University of Glamorgan (now South Wales). The absence of controlled vocabularies acted as a break on the impact of semantic technologies on interoperability and enabling data integration.

The development of Linked Open Data for cultural heritage is part of good practice helping to deliver Government Policy towards transparency and Open Data. The UK Government White Paper on Open Data (3) includes statements recognising that Public Data will be published in reusable, machine readable form under an open licence which enables free reuse, including commercial reuse to open standards following relevant recommendations of the World Wide Web Consortium. Moreover, Public Data from different departments about the same subject will be published in the same, standard formats and with the same definitions.

The SENESCHAL Project

The development of Linked Open Data for heritage requires investment in research and development time beyond the day-to-day resources of many organisations. Together with English Heritage, the Royal Commission on the Ancient and Historical Monuments of Wales (RCAHMW) and the Archaeology Data Service at the University of York, RCAHMS has been fortunate to work with domain experts, Doug Tudhope¹ as Principal Investigator and Ceri Binding² as Research Fellow, at the University of South Wales on the SENESCHAL project.

¹ http://hypermedia.research.southwales.ac.uk/people/tudhope/
² http://hypermedia.research.southwales.ac.uk/people/binding/
SENESCHAL, historically the steward or major-domo of a medieval great house (Concise Oxford English Dictionary), aims to provide Semantic ENrichment Enabling Sustainability of arCHAeological Links. It was a one year AHRC Knowledge Exchange project based on enhanced vocabulary services to make it significantly easier for vocabulary providers, such as RCAHMS, to make their vocabularies available as Linked Data and for users to index their data with uniquely identified (machine readable) controlled terminology that is semantically enriched and compatible with Linked Data.

**Project deliverables**

The project established [http://www.heritagedata.org/](http://www.heritagedata.org/) a brand neutral domain name for Cultural Heritage reference datasets. The website brings together multiple datasets (mostly thesauri), listed under the Scheme list

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>England</th>
<th>Scotland</th>
<th>Wales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monument Type</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Objects</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maritime Craft</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Period</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Event Type</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Archaeological Sciences</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Components</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building Materials</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evidence</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Cultural Heritage vocabularies available through the Heritagedata.org ([http://heritagedata.org/](http://heritagedata.org/)) website in February 2014.

For each scheme, there is an overarching definition of the dataset which expresses the properties of the concept in Linked Data format (Figure 2). These include Dublin Core terms defining the Dataset title (`dct:title`), description (`dct:description`) dates created (`dct:created`), modified (`dct:modified`) and issued (`dct:issued`) as well as the persistent URI for the dataset (`dct:identifier`) which is [http://purl.org/heritagedata/schemes/1](http://purl.org/heritagedata/schemes/1) for the Scottish Monument thesaurus. The attribution of the data provider is documented through the `dct:publisher`, `cc:attributionName` and `cc:attributionURL` properties. The licencing for each vocabulary is documented by the `cc:license`. Each contributor agreed to make their vocabularies available as Open Data. Datasets from England and Wales are available through a Creative Commons 3.0 licence

---

3 [http://heritagedata.org/live/getAllSchemes.php](http://heritagedata.org/live/getAllSchemes.php)
4 [http://dublincore.org/documents/2012/06/14/dcmi-terms/](http://dublincore.org/documents/2012/06/14/dcmi-terms/)
5 [http://creativecommons.org/licenses/by/3.0/](http://creativecommons.org/licenses/by/3.0/)
whereas the Scottish vocabularies are provided under an Open Government Licence\(^6\) to align with the attribution for our Canmore spatial data provided on terms analogous with an Ordnance Survey Open Data Licence\(^7\). For most of the vocabularies (excluding the maritime craft), the geographical coverage is defined by the Dublin Core term \texttt{dct:coverage} with the value provided from the Ordnance Survey Linked Data Dataset\(^8\). Finally \texttt{skos:TopConcept}\(^9\) documents the top level concepts associated with each concept scheme.

\(^6\) http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/
\(^7\) http://www.ordnancesurvey.co.uk/docs/licences/os-opendata-licence.pdf
\(^8\) http://data.ordnancesurvey.co.uk/datasets/os-linked-data
\(^9\) http://www.w3.org/2009/08/skos-reference/skos.html#hasTopConcept
Figure 2: Concept scheme for the Scottish Monuments Thesaurus (http://heritagedata.org/live/schemes/1)

The project established a URI for each concept within a scheme. Figure 3 shows the properties documenting the top level concept ‘COMMEMORATIVE’. The properties skos:topConceptOf and skos:inScheme define the vocabulary the concept belongs to whilst skos:narrower provides, potentially multiple, links to narrower terms related to the concept. The skos:preflabel indicates the preferred label for the concept, with skos:altLabel recording any alternative expressions of the concept whilst the Skos:scopeNote provides a definition of the Concept.

10 http://heritagedata.org/live/schemes/1/concepts/202.html
11 http://www.w3.org/2009/08/skos-reference/skos.html#topConceptOf
12 http://www.w3.org/2009/08/skos-reference/skos.html#inScheme
13 http://www.w3.org/2009/08/skos-reference/skos.html#narrower
14 http://www.w3.org/2009/08/skos-reference/skos.html#prefLabel
15 http://www.w3.org/2009/08/skos-reference/skos.html#scopeNote
One of the benefits of the Linked Data approach is demonstrated by the ability to express a Concept in Scots Gaelic, indicated by the gd suffix from the ISO 639-1 Codes for the representation of Languages as well as in English. Taking a term CENOTAPH\(^\text{16}\) from the thesaurus it is possible to map that concept ‘205’ to the preferred labels in English and Gaelic but also illustrate alternative labels for the Gaelic (Figure 4). The graph also visualises the relationship between the Concept and related (concepts 206\(^\text{17}\), 210\(^\text{18}\) and 1727\(^\text{19}\)) and broader terms (203\(^\text{20}\)) within the Scottish Monuments Thesaurus.

---

\(^{16}\) http://heritagedata.org/live/schemes/1/concepts/205.html  
\(^{17}\) http://heritagedata.org/live/schemes/1/concepts/206.html  
\(^{18}\) http://heritagedata.org/live/schemes/1/concepts/210.html  
\(^{19}\) http://heritagedata.org/live/schemes/1/concepts/1727.html  
\(^{20}\) http://heritagedata.org/live/schemes/1/concepts/203.html
Downloads, Services and Widgets

At the base of each record, there are links to enable users to download the vocabularies in various flavours of RDF (N-Triples, Turtle, JSON or XML). A series of REST URI calls\(^{21}\) have been developed for the vocabularies with results returned in a JSON structured string which permit AJAX callbacks for use in browser based applications. The project has also developed a suite of predefined visual user interface tools, or widgets\(^{22}\), published under a Creative Commons Attribution (CC- BY), to dynamically obtain information from web services provided from the controlled vocabularies. These tools provide vocabulary navigation, search, term suggestion and selection functionality that can be directly embedded within web pages.

Data alignment

Although datasets from the national vocabulary providers were already provided from controlled environments, much information from archaeological fieldwork is provided without reference to standardised terminologies. As part of the SENESCHAL project, a semi-automatic data alignment exercise, using the Levenshtein distance\(^{23}\)

\(^{21}\) http://www.heritagedata.org/blog/services/
\(^{22}\) http://www.heritagedata.org/blog/vocabularies-in-a-useful-form/
algorithm to define the minimal number of characters to transform one character string into another, was undertaken on information submitted through the OASIS\textsuperscript{24} (Online Access to Archaeological InvestigationS) form used by commercial archaeologists undertaking fieldwork through the planning process. Developed in 2003, before many of the thesauri were available, users index their fieldwork through a free text field often leading to typographical mistakes and other errors. These include simple spelling errors, alternate word forms (such as plurals), prefixes, suffixes and qualifiers (such as ‘possible’ used to convey uncertainty in identification) and detailed descriptive phrases and terms, such as ‘side wall of barn’ not covered by thesauri. The data alignment exercise has helped identify the inconsistencies to improve data quality for data provided by third parties.

A further exercise was undertaken to compare similar vocabularies from SENESCHAL partners using the same matching tool to generate output. Some of the vocabularies in use in Scotland and Wales were modelled on, and incorporated terms from, English Heritage thesauri but have subsequently developed separately. The 67.75\% exact match between the Scottish and English Monuments thesaurus reflects the common origin of the thesauri (Figure 5).

<table>
<thead>
<tr>
<th>Match</th>
<th>Percentage of matched terms in Scottish thesaurus</th>
<th>Example</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% match</td>
<td>67.75</td>
<td>Cord Rig (S)</td>
<td>Same concept although the Scope Notes differ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cord Rig (E)</td>
<td></td>
</tr>
<tr>
<td>90 to 99.99% match</td>
<td>0.92</td>
<td>Wheelhouse (S)</td>
<td>Different concepts:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wheel house (E)</td>
<td>A prehistoric dwelling (S) A building to house waterwheels (E)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Smokehouse (S)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Smoke House (E)</td>
<td>Same concept</td>
</tr>
<tr>
<td>80 to 89.89% match</td>
<td>3.43</td>
<td>Box Factory (S)</td>
<td>Different concepts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wax Factory (E)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alumina Works (S)</td>
<td>Same concept expressed differently</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aluminium Works (E)</td>
<td></td>
</tr>
<tr>
<td>Under 80% threshold / no match</td>
<td>26.9</td>
<td>Cruck Framed Building (S)</td>
<td>Same concept – expressed differently</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cruck House (E)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Colby Camp (S)</td>
<td>No concept for term in England</td>
</tr>
</tbody>
</table>

Figure 5: similarities and differences - data alignment between Scottish and English Monuments thesauri.

In comparing the character strings (to inform intellectual review), the data alignment exercise identified close matches (using a threshold of 80 \% and above) although manual judgement is then required to assess if the terms are the same or expressing different concepts. Those terms with less than 80\% match are either concepts expressed differently between the lists or terms not appearing in one or other thesaurus. The data alignment exercises are an important first step in formally aligning or amalgamating thesauri across Britain (and beyond).

\textsuperscript{24}http://oasis.ac.uk/pages/wiki/Main
Unlocking potential

If the data alignment exercise addresses retrospective problems from data supplied as free text, the widgets developed as part of SENESCHAL will help ensure consistency moving forward. Simply adding the widgets into the OASIS form during the planned redevelopment over the next couple of years, will ensure both consistency and currency in the terminology provided by the national heritage organisations with that available on the OASIS form and other online applications.

Adoption of persistent URIs for the terms used in heritage across national and local curators will help address issues of cross-searching datasets. For instance, PastMap (http://pastmap.org.uk) provides a map-based portal to historic environment datasets from Historic Scotland, RCAHMS and most local authority Historic Environment Records. Currently it is only possible to search by map location but by including concept IDs to ensure consistency between the respective datasets users will be able to search across data providers by term.

Inclusion of Scots Gaelic expressions of concepts demonstrates the potential for addressing multi-lingual searches and preserving regional and dialect expressions of terms through mapping to the concept.

SENESCHAL is an important first step in publishing cultural heritage data as Linked Open Data. The thesauri provide a major step towards interoperability, that is ensuring that data about the same subject will be published in the same, standard formats and with the same definitions.

Acknowledgements

The SENESCHAL project was supported by the Arts and Humanities Research Council [grant number AH/K002112/1]. The project was led by Douglas Tudhope and Ceri Binding at the University of South Wales and I am very grateful for their helpful input to this article. I would also like to thank Kirsty Lingstadt and Robin Turner at RCAHMS for their comments on the paper.

References


http://gtr.rcuk.ac.uk/project/C77A721A-6427-4EE0-BC46-07899B933D44
In June 2011, the RDA Test Coordinating Committee recommended that the US national libraries should implement RDA but subject to certain conditions, one of which was that they should “Demonstrate credible progress towards a replacement for MARC”. The Committee found that those trying out RDA “felt any benefits of RDA would be largely unrealized in a MARC environment”. This is not the first time that MARC has been seen to be in need of replacement. Some twelve years ago, Roy Tennant wrote a now often quoted article in Library Journal entitled *Marc Must Die*:

> The problems with MARC are serious and extensive, which is why a number of us are increasingly convinced that MARC has outlived its usefulness.

This article will seek to address some of these problems, including most of those mentioned by Tennant in his article. However, when talking about MARC it is worth bearing in mind Tennant’s words on the difficulties of isolating MARC:

> When I refer to MARC in this column, I am conflating several interrelated things. There are the MARC syntax, the MARC data elements, and the Anglo-American Cataloguing Rules (AACR).

To this must be added RDA. This article will refer to MARC21, which is the dominant, if still not universal standard. Some problems, particularly those to do with punctuation, are less pronounced in UKMARC, for example. To many non-cataloguers, including those involved in systems work, MARC is cataloguing. The existence of the underlying content standards- AACR2 and RDA- is something of a mystery and, given the closed nature of them both, not a mystery that is easily solved. Indeed, it is the need for systems to reach beyond the cataloguing unit, the OPAC, and the library itself that generate much of the need to reassess MARC. Even before the Bibliographic Framework Initiative proposed a linked data approach the RDA Test Coordinating Committee already recognised that the problems with MARC stood in the way of translating RDA data to a linked data environment: “MARC may hinder the separation of elements and ability to use URIs in a linked data environment.”

**Brief History**

It is worth reflecting briefly on how MARC came to be as it is. Before MARC, catalogue records were generally recorded in print form- e.g. an index card or dictionary catalogue- or perhaps on a microfilm.
In either case, the point is that the record is designed to be read by eye by humans. The catalogue card is essentially text to be read in sequence and as a whole. The pre-AACR2 example above uses layout and some punctuation to separate out elements of data. Extracting out a line of data “London, 1965” only makes sense to us as we recognise “London” as a place and “1965” as a date. However it still needs the rest of the card for us to know which book was published in London in 1965.

MARC’s basic function is to encode catalogue data, such as AACR2, onto a computer. By itself, a raw MARC record is essentially unreadable:

```
00788nam a2200181 a 450000100270000005001700027008004100044024001500085245021000100260004900310 3000332003595040041003916500033004327000023004657100039004887100030005277100 04900557 UCL01000000000000000477125_20061112120300.0 850710s1965__00_aModels for decision : ba conference under the auspices of the United Kingdom Automation Council organised by the British Computer Society and the Operational Research Society / edited by C.M. Berners-Lee. ; bEnglish ; c Universities Press, c1965. ; dax, 149 p. ; bill. ; c23 cm. ; a Includes bibliographical references. ; 0 a Decision making ; v Congresses. 1 a Berners-Lee, C. M. 2 a United Kingdom Automation Council. 2 a British Computer Society. 2 a Operational Research Society. (Great Britain). ""`
```

In the above example, anything to do with the 245 field is marked in bold. You will notice that the 245 field tag is only mentioned in what is called the directory (the shorter bold sequence); the indicators and text of the field are carried separately. The subfield codes are marked with usually hidden characters, which have been changed to underscores above to make them visible. To extract any piece of data from this record, the whole record has to be disassembled, something that requires special software such as a Library Management System (LMS) or Marcedit.

A MARC display modified for human consumption, such as most of us are used to when editing MARC records or using MarcEdit, is more readable for cataloguers at least, although this is no longer a standard MARC record:

<table>
<thead>
<tr>
<th>Field</th>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>245 00</td>
<td>$a</td>
<td>Models for decision:</td>
</tr>
<tr>
<td></td>
<td>$b</td>
<td>a conference under the auspices of the United Kingdom Automation Council organised by the British Computer Society and the Operational Research Society /</td>
</tr>
<tr>
<td></td>
<td>$c</td>
<td>edited by C.M. Berners-Lee.</td>
</tr>
<tr>
<td>260__</td>
<td>$a</td>
<td>London:</td>
</tr>
<tr>
<td></td>
<td>$b</td>
<td>English Universities Press,</td>
</tr>
<tr>
<td></td>
<td>$c</td>
<td>1965.</td>
</tr>
<tr>
<td>300__</td>
<td>$a</td>
<td>x, 149 p.:</td>
</tr>
<tr>
<td></td>
<td>$b</td>
<td>ill.;</td>
</tr>
<tr>
<td></td>
<td>$c</td>
<td>23 cm.</td>
</tr>
<tr>
<td>504__</td>
<td>$a</td>
<td>Includes bibliographical references.</td>
</tr>
<tr>
<td>700 1_</td>
<td>$a</td>
<td>Berners-Lee, C. M.</td>
</tr>
</tbody>
</table>

The MARC record, in this case including all the ISBD punctuation, is ideal for carrying the index card’s data for human consumption, whether on index cards, compiling a dictionary catalogue, or even a display screen for an OPAC, at least if you want to recreate the index card layout. The rest of this article will concentrate on some specific problems with MARC. In these cases, it helps to think of oneself as a computer or a computer programmer. One way to do this is to imagine writing out specific and foolproof instructions for finding specific pieces of information for someone who has no idea of the subtleties of MARC or cataloguing rules: how complex do they become and how many exceptions do you need to put in?

**Finite Notation Problem**

MARC uses field names that are three figure numbers, two numerical indicators, and single letter subfield codes. Like the Dewey Decimal (DDC) system, the notation itself can become constricting. For example, using the second indicator of the 650 field for the subject scheme…

<table>
<thead>
<tr>
<th>Field</th>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>650 _0</td>
<td>for LCSH</td>
<td></td>
</tr>
<tr>
<td>650 _1</td>
<td>for LC for Childrens</td>
<td></td>
</tr>
<tr>
<td>650 _2</td>
<td>for MeSH</td>
<td></td>
</tr>
</tbody>
</table>

…means that that there are only ten possible schemes that can be recorded. In the same way as DDC uses 0 to expand, MARC in this case resorts to a second indicator 7 and an additional subfield:

<table>
<thead>
<tr>
<th>Field</th>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>650 _7</td>
<td>Source specified in subfield $2</td>
<td></td>
</tr>
</tbody>
</table>

This solves the immediate problem while creating a degree of inelegance and forcing any system to look in two places. Another example that may be familiar is that of the 246 field for variant title. The first indicator controls the addition of a note or added entry, the second indicator details what kind of varying title is involved; if the title begins with an article, there is no way to show this as there is with the second indicator of the 240, 245, 730, or 740. There needs to be a third indicator and there is no way to add one:
Data in More than One Place

The solution to the limited number of subject schemes in the previous section forced data to be in more than one place. This is common in MARC. For instance, on an index card, this might be recorded in human-readable form in a textual note:

546 __ $a In English.

A computer readable form is provided in positions 35-37 of the MARC21 008 field where it would be recorded using a code: “eng”. If the language of an item is more complicated- e.g. for a multilingual book- then these three characters are not sufficient and we run into the finite notation problem again, and the 040 field is used:

041 __ $a eng
   $a ger
546 __ $a English and German parallel texts.

Furthermore, languages can appear in other places, such as part of uniform titles, this time not free text, but not using the same codes as the 008 or 041 either:

240 10 $l English

A computer programme trying to work out what language(s) a book was in or had been translated from could end up having to look in at least three MARC fields and understand two coding systems; the arguably richest field- the 546- would be effectively lost as it is too difficult to process English text (assuming the catalogue record is in English!) and the many possible wordings cataloguers may have used.

Double Encoding

One notable feature of MARC21, especially as compared with something like UKMARC, is that the data is effectively encoded twice: first using ISBD punctuation as mandated by AACR2, then again using MARC21 fields and subfields. RDA in theory only defines elements and is free from ISBD, but to enter RDA into MARC21 means putting the ISBD punctuation in. The following example uses ISBD punctuation only, i.e. no MARC:


As pure text, this is confusing for the uninitiated and looks like biographical information about the composer Edward Elgar. In the context of a catalogue record display and in the right place, I know by the colon and the comma that “Worcester” is a place of publication, “Edward Elgar” is in fact a publisher, and “1857-1934” is the range of publication dates. Similarly, the coding alone of this 260 field makes these elements clear:

260 __ $a Worcester
   $b Edward Elgar
   $c 1857-1934.

Putting them both together is unnecessary:

260 __ $a Worcester :
   $b Edward Elgar,
   $c 1857-1934.
It makes public display on a traditional OPAC easier if the intention is to replicate the catalogue card, but this can cause significant problems if you want to isolate a piece of data.

**Text, Not Data**

For instance, what is the place of publication of the book in the foregoing example? In MARC, this should be a straightforward as the place of publication is recorded in the 260 field (at least in AACR2) subfield $a. This means the place of publication is:

Worcester :

This is clearly not the case as the place is actually:

Worcester

A computer programme wanting to display this information separately on a catalogue, or a list of places, or to index it, has to strip the space-colon away. If the record doesn't have a publisher or date there might even be commas or full-stops instead. If it’s a cataloguer-supplied piece of data, there might also be square brackets to dispense with, so a computer programme now has to look out for any of the following to make sense of this information:

Worcester :
Worcester,
Worcester.
[Worcester] :
[Worcester],
[Worcester].

When that is done, the information is still not easy for a computer to understand in that “Worcester” is a human-readable, if standard English, string. However, under RDA, not counting spelling or transcription errors, this could also potentially be:

Worcester, Worcestershire
Worcester, Worcs
Worcester, Worcestershire, England
Worcester, England

(Not to mention the ones in South Africa, Massachusetts, New York, and elsewhere in the U.S.). Automatically establishing that these are the same thing is difficult, which is where a linked data approach is helpful. A URI such as [http://id.loc.gov/authorities/names/n81018214](http://id.loc.gov/authorities/names/n81018214) identifies the place as Worcester (in England) regardless of the transcribed form.

MARC stores a number of other elements that should be data as text instead. Sizes of books are stored with their units, and the units may vary; the 264 _4 subfield $c should store a copyright date but instead holds a string starting with a copyright symbol: to get the actual date, the copyright symbol needs to be removed.
Data Mixed Up

Different types of MARC data are frequently mixed up and take some disentangling. Consider the following:

245 00 $a Data on the web
$h [electronic resource] :
$b research and applications / 
$c Antonis Bikakis, Adrian Giorca (eds.).

The GMD in the 245 subfield $h is not a part of the title except that AACR2 deals with it as such. If a library chose not to display the GMD- perhaps preferring icons instead- then removing the $h is not as simple as it could be as punctuation is removed at the same time:

245 00 $a Data on the web
$b research and applications / 
$c Antonis Bikakis, Adrian Giorca (eds.).

The 245 subfield $c is defined in MARC as the “Statement of responsibility, etc. “. In the following example, the “etc.” includes two parallel statements of responsibility, a title proper, and a subtitle:

245 10 $a Enduring resistance :
$b cultural theory after Derrida / 
$c edited by Sjef Houppermans, Rico Sneller, Peter van Zilfhout. = La résistance persévère : la théorie de la culture (d’)après Derrida / édité par Sjef Houppermans, Rico Sneller, Peter van Zilfhout.

It is still useful for displaying a catalogue record, but it is now hard to isolate any statements of responsibility, not to mention the parallel title, which is essentially lost. The 245 $c effectively loses any meaning.

Changing Text As Primary Key

Current authority control practice using MARC requires a match between the form of the heading in a bibliographic record and the corresponding authority record. This requires considerable maintenance work and potential for broken links. The heading for Oscar Niemeyer was previously:

Niemeyer, Oscar, 1907-

But changed on his death to:

Niemeyer, Oscar, 1907-2012.

This requires both the authority and bibliographical records to be changed. With record sharing and outsourcing of either type of record, this effectively needs changing on all systems at all libraries. When computers are doing the indexing and matching, even small details can break links between records:

Niemeyer, Oscar, 1907-2012.
Niemeyer, Oscar, 1907-2012

Using URIs such as http://id.loc.gov/authorities/names/n82013357 for Niemeyer avoids such problems. Forms of name, death dates, or any other information can be added or changed without breaking the link itself.
Expressing Relationships

RDA has at least improved MARC’s handling of relationships. In the following example, it is impossible to know, at least in isolation, what C.M. Berners-Lee’s role is with regard to the item being catalogued:

700 1_ $a Berners-Lee, C. M.

He could be a second author, an editor, a translator, or perhaps even the subject of a Festschrift. Relationship designators will help to ease this:

700 1_ $a Berners-Lee, C. M.,
  $e editor of compilation.

Although, note that the comma before the $e is another defiantly textual addition. Remove the $e as we did with the GMD above, and we are left with this:

700 1_ $a Berners-Lee, C. M.,

The comma now looks like part of the name!

Handling of RDA

Of course, what has really accelerated the proposed demise of MARC is the implementation of RDA. Some of these issues, such as the greater importance of relationships in RDA have already been mentioned. MARC is a record-based format. This makes it difficult to disentangle data for the Manifestation, Expression, and Work elements of an RDA description. Implementation of RDA along the lines of the Joint Steering Committee (JSC)’s Scenario 1 would require separate records for each WEMI entity. The MARC record is based on the Manifestation. Expression-level data is found mixed in the MARC record. In the following simple RDA record, the Expression-level data is marked in bold:

245 00 $a Models for decision :
  $b a conference under the auspices of the United Kingdom Automation Council organised by
  the British Computer Society and the Operational Research Society /
  $c edited by C.M. Berners-Lee.
264 _1 $a London :
  $b The English Universities Press Limited,
  $c 1965.
300 __ $a x, 149 pages :
  $b illustrations ;
  $c 23 cm
336 __ $a text
  $2 rdacontent
337 __ $a unmediated
  $2 rdamedia
338 __ $a volume
  $2 rdacarrier
504 __ $a Includes bibliographical references.
700 1_ $a Berners-Lee, C. M.,
  $e editor of compilation.

---

It is difficult to express any relationships between the Expression and Manifestation here. Some data is even mixed up within MARC fields: in this case the illustrative material in 300 subfield $b$. Some fields, such as the 700, are not even identifiable as any Work or Expression elements until they are followed by a relationship designator: "author” would make it a Work element, “editor of compilation” shows it is an Expression element.

Who uses MARC?

Only libraries use MARC. When MARC’s use is restricted to creating index cards or compiling a catalogue for internal use, this might be fine, but it imposes restrictions on how libraries can interact with other organisations. By continuing to use MARC, libraries are tied to library-specific systems and processes. For instance, it would be difficult to move beyond a library management system for metadata management and editing, whereas there are tools for dealing with standards like RDF which are not so restricted to libraries. The restriction of MARC to the library community also makes it harder for non-librarians to use and benefit from the rich data libraries hold and continue to create.

Even within the library, MARC is by no means universal. It is as noteworthy that repository databases are not using MARC as it is that AACR2 was not the obvious method of recording repository metadata. Discovery systems like Primo also bring in a great range of metadata to work alongside the catalogue’s MARC records; MARC records in Primo are also converted to an internal format for display and indexing.

Conclusion: the Purpose of MARC

If MARC is not used for display- and arguably it hasn’t for some time given the line-by-line display favoured by online catalogues over the card index display- then what is its purpose? Cataloguers interact with MARC most often through input interfaces. Why this happens, though, is not obvious. The rules we have used for years- AACR2 and to a greater extent RDA- have elements which are then encoded in MARC fields and subfields. For example, the title proper is entered into the 245 subfield $a$; the publisher in 260 subfield $b$. Why is the cataloguer faced with an extra layer of numbers and codes to learn, especially as they are then expected to impose ISBD punctuation as well? It is perfectly possible to create a system which asks the user to fill in these elements instead of MARC codes. We use systems like this all the time on the web in various web forms where the underlying structure of the database is hidden from us. A practical example is the software developed by the company The MARC of Quality called RIMMF (RDA in Many Metadata Formats)\(^7\). This allows the cataloguer to input RDA elements directly without using MARC or ISBD, as in the following screenshot:

---

\(^7\)The MARC of Quality. RIMMF2. http://www.marcofquality.com/
Here the cataloguer can concentrate on applying the rules and structure of the data. It is not necessary at the cataloguing stage to know how this is stored internally - in a relational database, raw MARC files, a linked data triplestore, or some other means. One reason why MARC input displays may have persisted is that, as Rob Styles once remarked, MARC has become the "lingua franca of library cataloguing".

We have already discussed how MARC, especially as marked up with ISBD, can make manipulation of data difficult. For instance, indexing authors with relationship designators becomes more tricky if you have to worry about the comma that may be present (but isn’t always) before a subfield $e. Terry Reese’s MarcEdit software allows cataloguers to manipulate MARC records by converting them to a text format that is easy to deal with, i.e. that isn’t MARC. It converts them back so the library system can understand them again.

The real strength of MARC is as a medium of exchange and distribution. It is compact and all major library management systems understand it. There are systems and standards, such as z39.50, that make exchanging records straightforward. However, the system works mostly on the principle of search and delivery of records. Linked data offers the possibility of publishing data more openly for wider consumption using HTTP.

MARC has been a very successful, long-lasting, and efficient servant of library data, getting library catalogue data quickly onto computer systems for local use, distribution, and sharing. However, technology standards, and even cataloguers’ own rules, are moving beyond what MARC is capable of. Its narrow library focus, record-based structure, and reliance on text-based cataloguing are limiting what libraries can do with the rich and high quality bibliographic data cataloguers have been amassing for decades and continue to produce.

---

8 Carty, Celine. RIMMF, or, Cataloguing Without MARC. Slide 7.
References


Bibframe is an initiative of the Library of Congress (LC) to provide a means of expressing bibliographic data for the future. One of the main motivations for embarking on the Bibframe project is to replace MARC. The need to replace MARC was largely born out of the U.S. national libraries' testing of RDA:

Many survey respondents expressed doubt that RDA changes would yield significant benefits without a change to the underlying MARC carrier.\(^1\)

Bibframe is short for the BIBliographic FRAMEwork Initiative. It is being developed by LC together with the consultants Zepheira and a number of partners or early experimenters. Zepheira are consultants with expertise and experience with the Semantic Web, including work with OCLC on Schema.org as well as a new project with University of California, Davis to "investigate the future of research library operations, particularly the production of metadata — or data on data — and deployment on the Web." Its president, Eric Miller, has been prominent in the development of the Semantic Web and RDF. The partners include the British Library, Deutsche Nationalbibliothek, George Washington University, National Library of Medicine, OCLC, and Princeton University.

This article will focus primarily on the Bibframe model itself and what it looks like. A suggestion is also made for how cataloguers can compare existing MARC examples with how they might be represented in Bibframe, although care must be taken considering that Bibframe is still very much under development and conversions are automatic.

Terminology and Examples

It is worth dwelling at least briefly on the name itself, which is officially all in capitals (BIBFRAME), although it is not an acronym. It seems, perhaps thankfully, to be not unacceptable to spell it with an initial capital only. In the interests of calm, Bibframe will be used throughout this article.

Examples will generally use the Turtle serialization of RDF. In brief, RDF represents all data as triples, which are simple statements with three elements:

Subject - Predicate - Object

The Subject and Object are entities (thing, person, concept, anything) related by the Predicate. So, the Subject might be a book, the Object an author, related by a creation relationship as the Predicate. Each of these is normally identified by a URI— basically a URL used as an identifier, or a string of text. For example, http://id.loc.gov/authorities/names/n78095332 is a URI coined by LC for Shakespeare. These can be shortened for readability by using a system of prefixes. The URI for the Bibframe creator relationship is http://bibframe.org/vocab/creator. With the following line…

```
@prefix bf: <http://bibframe.org/vocab/> .
```

…this can be represented instead more succinctly as `bf:creator`. The prefix `bf:` will be used throughout this article although the prefix declaration will be omitted in examples; the prefix `ex:` will be used to refer to made up entities used as examples. Note that where several triples have the same Subject, the Subject is not repeated on second and subsequent lines. Generally each triple ends in a full stop or a semicolon.

---


The Bibframe Model
The basic Bibframe model is illustrated below.

Those familiar with the FRBR model will notice some similarities, in particular entities, or things, linked together with relationships. There are Works and Instances representing FRBR Group 1 entities, and Authorities representing Group 2 and 3 entities. There are also obvious differences such as the Group 1 entities being separated into Works and Instances rather than Works, Expressions, and Manifestations. We will consider these types of entity in turn.

Work (and Expression...)

The Bibframe Work is “a resource reflecting a conceptual essence of the cataloging resource”. This is clearly similar to the FRBR Work defined as a “distinct intellectual or artistic creation”\(^4\) The FRBR Expression- “the specific intellectual or artistic form that a work takes each time it is ’realized’”\(^5\) is also conceptual: a specific French translation of Evelyn Waugh's *Decline and Fall* is still an abstract idea not tied to a particular printing or copy.

---


Whereas Bibframe Instances map fairly well to FRBR Manifestations, there is an absence of anything where one might expect a FRBR Expression which is also conceptual: “the specific intellectual or artistic form that a work takes each time it is 'realized'”. A specific French translation of Evelyn Waugh's *Decline and Fall* is still an abstract idea not tied to a particular printing or copy. However, Bibframe is not intended to be tied tightly to RDA and FRBR in the same way that MARC was tied to AACR; Bibframe is intended to accommodate a wider range of metadata and be useful to a wider community than traditional cataloguers:

“In addition to being a replacement for MARC, BIBFRAME serves as a general model for expressing and connecting bibliographic data.”

It can still accommodate FRBR Works and Expressions and some attempt will be made in the following account to relate the two models.

That Bibframe, FRBR, and RDA all use the same word “Work” does not necessarily mean they are the same thing. None of these are necessarily the same thing as the WorldCat Work described by Richard Wallis as:

a high-level description of a resource, containing information such as author, name, descriptions, subjects etc., common to all editions of the work

This is itself based on the CreativeWork defined by the Schema.org vocabulary. However, they clearly have some similarity and the Bibframe Work can potentially accommodate these and other models. So, a FRBR Expression could also be represented as a Bibframe Work. How could this work in practice? First we will examine some basic properties of a Bibframe Work.

A Bibframe Work is a type of Bibframe Resource. A Resource in this sense is the linked data sense of anything that can be given a name and described. A number of properties can be applied to any Resource, e.g.

- bf:authorizedAccessPoint
- bf:identifier
- bf:label
- bf:relatedTo

---


7 To put it more flippantly: “Bibframe has worked on modelling works as Works within the Bibframe model, similar to the RDA modelling work, itself modelled on the work on the FRBR model of Works and Expressions. A Bibframe Work is a creative work, perhaps a FRBR Work, or an RDA FRBR Work but it also expresses a FRBR Expression, and of course an RDA FRBR Expression. A Work may express another Work based on others’ work, not just a FRBR Work or an RDA Work. That also works. FRBR Works or RDA Works expressed as Bibframe Works can relate to FRBR Expressions (Bibframe Works or RDA Expressions). So, Works are works that can be Works but also Expressions linked to Works that really are Works.” Meehan, Thomas. The BIBFRAME Work. http://www.aurochs.org/aurlog/2013/05/25/the-bibframe-work/


An **identifier** is a “number or code that uniquely identifies an entity”; a **label** is most commonly the textual name of something: “Bombus vestalis”, “Tom”, “Evelyn Waugh”, “Decline and fall”; an **authorizedAccessPoint** access point would be a controlled name such as we are used to in authority work. All of these can be applied to a Work. In addition, Bibframe lays out a great many additional properties specifically for a Work, including the following:

- bf:contributor
- bf:creator
- bf:language
- bf:title

It would be straightforward to imagine how relevant RDA elements could be mapped to these (albeit with some interesting questions about subtitles, part titles, parallel titles and so on...). Note however that contributor and language are RDA Expression elements, while creator and title are RDA Work elements. There are also properties showing relationships to other Bibframe Works, for example:

- bf:translation
- bf:translationOf

In the above case, these would be found in RDA Expression records. What really clarifies this are the following two properties of Bibframe Work:

- bf:expressionOf
- bf:hasExpression

In the following example, eg:wk0123 represents Evelyn Waugh's Brideshead revisited, an RDA work, as a Bibframe Work:

```
ex:wk0123 a bf:Work ;
  bf:hasExpression ex:exp0456
```

The first line says that eg:wk0123 is a Bibframe Work; the second gives a controlled name, in this case the LC authorized form; the third line says that this FRBR Work is related to a FRBR Expression. Here is an example for that Expression:

```
ex:exp0456 a bf:Work ;
  bf:hasExpressionOf ex:wk0123 .
```

The first line says that ex:exp0456 is also a Bibframe Work; the second again gives a controlled name, in this case qualified by language; the third line relates the FRBR Expression back to the FRBR Work. So, ex:exp0456 is both a Bibframe Work and a FRBR Expression.

**Instance**

The Bibframe Instance is distinct and is analogous to the FRBR Manifestation: a “resource reflecting an individual, material embodiment of the Work.” Any of the four Resource properties can be applied to the Instance, as well as a number of specific ones, including:

- bf:edition
- bf:isbn
- bf:instanceOf
- bf:publication
- bf:titleStatement

---

Edition and isbn are straightforward; titleStatement is for a transcribed title, such as you might find in AACR2 or RDA, and publication links to further details about places, publishers, and dates. InstanceOf provides a link back to a Bibframe Work, so we might extend our example as follows, also adding the complementary Work property hasInstance:

```
ex:wk0123 a bf:Work;
  bf:hasExpression ex:exp0456.

ex:exp0456 a bf:Expression;
  bf:expressionOf eg:wk0123;
  bf:hasInstance ex:inst0789.

ex:inst0789 a bf:Instance;
  bf:titleStatement "Vozvrashchenie v Braĭdshkhed";
  bf:instanceOf ex:exp0456.
```

In this example, ex:wk0123 is a FRBR Work represented as a Bibframe Work; ex:exp0456 is a FRBR Expression also represented as a Bibframe Work; finally, ex:inst0789 is a FRBR Manifestation represented as a Bibframe Instance. They are related to each other and have properties of their own.

**Authority**

The Bibframe Authority represents “People, Places, Topics, Organizations, etc.” However, the way it does so is not necessarily quite so straightforward as one might expect in a linked data context. The classic linked data example of a book and its author looks something like this:

```
Or, in Bibframe triples:

ex:wk666 a bf:Work;
  bf:creator <http://id.loc.gov/authorities/names/n79049248>.
```

The Work ex:wk666 has a creator identified by the LC URI on the right, which represents Evelyn Waugh. However, Bibframe is concerned to preserve some aspects of traditional authority control. A typical name index on a library catalogue might be made up of authorised headings linked to locally held LC Authority records as well as unauthorised headings with no corresponding authority record and therefore, in linked data terms, no URI to link to. The library might not undertake its own authority work or might simply have a large number of headings it does not have the time to authorise. Bibframe uses what it calls a "lightweight abstraction layer" to sit in the middle between a Work or Instance and an external authority:

---

In Bibframe triples, the arrangement looks like this:

```turtle
ex:wk666 a Work ;
  bf:creator ex:person99
ex:person99 a bf:Person ;
  authorizedAccessPoint "Waugh, Evelyn, 1903-1966."
  hasAuthority <http://id.loc.gov/authorities/names/n79049248>
```

In this example, the book again has a creator, but they are not identified directly by the LC Authorities URI. Instead, a local authority - `ex:person99` - is used: this is a Bibframe Authority. There are several different kinds of Bibframe Authority: agents, places, times, and topics; in turn, there are several types of Agent: Person, Family, Organization, Jurisdiction, and Meeting. Evelyn Waugh is clearly a Person, so the first triple of the Authority section says so:

```turtle
ex:person99 a bf:Person ;
```

We can assign it a heading using the `authorizedAccessPoint` property:

```turtle
```

This is the same as the 100 field in a MARC authority record which records the definitive textual string identifying a person. We can give the specific URI for this authority so the system has a link to follow, or as a source of variant or updated names:

```turtle
ex:person hasAuthority <http://id.loc.gov/authorities/names/n79049248> .
```

This explicit linking is not something we can currently do in MARC without following something along the lines of Karen Coyle's recommendation to use the subfield $0^{14}$. Bibframe also provides ways of recording which authority scheme has been used, e.g. Library of Congress or MeSH, using the `authoritySource` property (akin to the second indicator of a 650 field).

**Annotation**

Annotation is the fourth class of resource in the Bibframe model which adds further information about a resource such as a Bibframe Work\(^{15}\). An Annotation could comprise cover art, a review, summary, or, more controversially, holdings. In these cases, Annotations also provides a mechanism to record who made the link between, for example, a book and a review of the book. Individual libraries might for example want to highlight different book reviews for their readers or display cover art matching the actual printing they have in stock. Below is an example in Turtle for linking the Work *Brideshead Revisited* with a summary of the book:

---


\(^{15}\) Bibliographic Framework Initiative. *Annotation*. http://bibframe.org/vocab/Annotation.html
The first two triples say that the book ex:wk005 is a Bibframe Work and that it is an edition of *Brideshead Revisited*. The third triple make the link to the Annotation. The Annotation named ex:ann010 in this case identifies itself as a Bibframe Summary and makes a reciprocal link back to the Work using the the summaryOf property. The annotationAssertedBy property is the part which specifies which library wanted to link the book to the summary; the URI above is from the list of LC institution codes (also available as linked data)\(^\text{16}\) and specifies University College London (UCL). The annotationSource gives the source of the summary, in this case DBpedia (the linked data version of Wikipedia)\(^\text{17}\) and the startOfSummary contains the start of the summary itself. Bibframe provides a number of similar properties for the whole summary or other type of Annotation.

**Getting More Examples**

Examples are often the best way to get a grip on this kind of thing and the Bibframe website provides a means of getting Bibframe data that has been converted from LC's MARC21 record.

- Go to the [MARC to BIBFRAME Comparison Service](http://bibframe.org/tools/compare/)
- Enter an LC system number (e.g. 10342843) and click on Run Comparison
- Select BIBFRAME RDF/XML view. As RDF/XML is not always easy to read, you might want to convert this to Turtle, as used in this article, in which case:
  - Copy the result and paste it into the Input Field tab at this RDF converter [http://rdf-translator.appspot.com/](http://rdf-translator.appspot.com/). Make sure there are no stray blank lines or bits of text.
  - Select N3 (i.e. Turtle) output.
- Submit!

**The Future**

Bibframe is still very much a work in progress. The general model described above has been fairly stable since it was first proposed although the precise properties and details have changed considerably over time. It will doubtless continue to do so although in 2014 the emphasis moved more towards testing the existing work. An “Implementation Testbed” has been set up. This is aimed at organisations who will be expected to:

- have developed or be developing a BIBFRAME implementation;
- participate in testing;
- participate in listserv discussion;
- report results of testing;
as well as “make an earnest effort to participate in conference calls” 18. LC is also maintaining a “BIBFRAME Implementation Register” of planned or existing implementations. Activities by several of the partners and Colorado College have already been registered. 19 One of the most interesting is the Deutsch Nationalbibliothek’s addition of an option to view a Bibframe representation of a record on their catalogue. 20

The big question is of course when, and indeed if, Bibframe will be widely adopted. It is certainly impossible to predict at this point. It has the backing of LC as well as significant libraries as partners, but there is clearly a shift needed in either the capabilities of library management systems or a change to the systems used by libraries before adoption can be contemplated if successful adoption means the replacement of MARC. There is no Scenario 3 where we can use existing systems with a few tweaks. This applies of course to any linked data implementation and, in some respects, Bibframe is already playing catch-up. The British National Bibliography, for instance, already publishes linked data using its own data model. 21 Like the BNB, the Europeana data model re-uses elements coming from already-established vocabularies, such as Dublin Core, OAI-ORE, SKOS and CIDOC-CRM, thus lowering the cost of its creation and, hopefully, its adoption 22.

This is in contrast to the approach taken by Bibframe:

There are many benefits of vocabulary reuse, but as with many things, there are costs as well that need to be carefully considered. Designing systems that leverage multiple vocabularies managed by various stakeholders is a tricky issue and one that requires careful consideration. There are many reasons why namespaces/vocabularies "drift" over time ("not found" errors being a worse case example) and all of these may have an affect on systems. Business acquisitions, economic factors, organizational changes, changing social interests, etc. are just a handful of reasons for causing such change. Thinking ahead to infrastructure to support the next 40+ years of libraries, namespace persistence is a key point to consider when dealing with how best to integrate and invest in vocabulary terms outside of ones community 23.

With RLUK joining the European Library in 2013 24 there is also the prospect of 200 million records from UK academic libraries becoming available as linked data using the Europeana Data Model. Will there be incentives for this data to be converted to Bibframe, whether as a replacement or in parallel, especially when the re-use of existing vocabularies is often considered as more within the spirit of linked data?

21 British Library. Free Data Services: Linked Open BNB. http://www.bl.uk/bibliographic/datafree.html#lod
It is perhaps likely that there will not be a direct substitution of Bibframe for MARC. It seems unlikely, and undesirable, that cataloguers will be inputting bibliographic data using Bibframe vocabulary directly, although the continued prevalence of MARC-based input forms might cast some doubt on that. There has been some work on developing a prototype Bibframe editor, as described at an ALCTS forum in January 2014:

A demo of BIBFRAME Editor illustrated a dynamic interface, in which users select an item type (such as image, electronic article or paperback book), and enter information into the BIBFRAME Editor form. As the user enters properties, like subject, artist, author, or language, a dropdown list appears and the user can choose the desired description, in a way that is visually the same as what we have become used to seeing as we type search terms in a search engine. Images, such as cover art for an album, can be dragged and dropped into the BIBFRAME Editor as well. Once completed, that record can be saved and exported.

The move from MARC to a linked data solution will in many ways be more profound than that from AARC2 to RDA. To what extent Bibframe is the principal answer, or instead part of a largely hybrid solution in partnership with other models and vocabularies, remains to be seen.

References


British Library. Free Data Services: Linked Open BNB. http://www.bl.uk/bibliographic/datafree.html#lod


DBpedia. http://dbpedia.org/About


---


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
Catalogue & Index is electronically published by the Cataloguing and Indexing Group of the Chartered Institute of Library and Information Professionals (CILIP) (Charity No. 313014)

Subscription rates: free to members of CIG; GBP 15.00 for non-members & institutions

Advertising rates: GBP 70.00 full-page; GBP 40.00 half-page. Prices quoted without VAT.

Submissions: In the first instance, please contact the Editor:

For book reviews, please contact the Book Reviews Editor: Neil T. Nicholson, Cataloguing & Metadata Services Team Leader, National Library of Scotland, e: n.nicholson@nls.uk

ISSN 0008-7629

CIG website: http://www.cilip.org.uk/cataloguing-and-indexing-group/catalogue-index

CIG blog: http://catandindexgroup.wordpress.com/

CIG twitter @CILIPCIG

Tags from the CIG blog on specific areas of interest:

authority control book reviews Catalogue and Index cataloguing CIG activities CIGS classification committees and working groups conferences Dewey digitised material Dublin Core events folksonomies linkblog MARC metadata news RDA Semantic Web social software standards taxonomies UDC