Publishing the British National Bibliography as Linked Open Data

Corine Deliot
Metadata Standards Analyst
The British Library

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.

---

[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. It was originally published in *Catalogue & Index* March, 2014, Issue 174, pages 13-18 and has undergone a very minor edit to update the link to our SPARQL endpoint.

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.

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[2] 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" [3]. Finally, we had to consider how to produce valid URIs, i.e. conformant with the URI syntax specified by IETF[4] [5].

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. [6] 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[7] 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.

---

[2] VIAF: The Virtual International Authority File http://viaf.org/

[3] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60975/designing-URI-sets-uk-public-sector.pdf.

[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

They also provided granularity, thus enabling us to avoid mapping all of these to a generic dcterms:description.

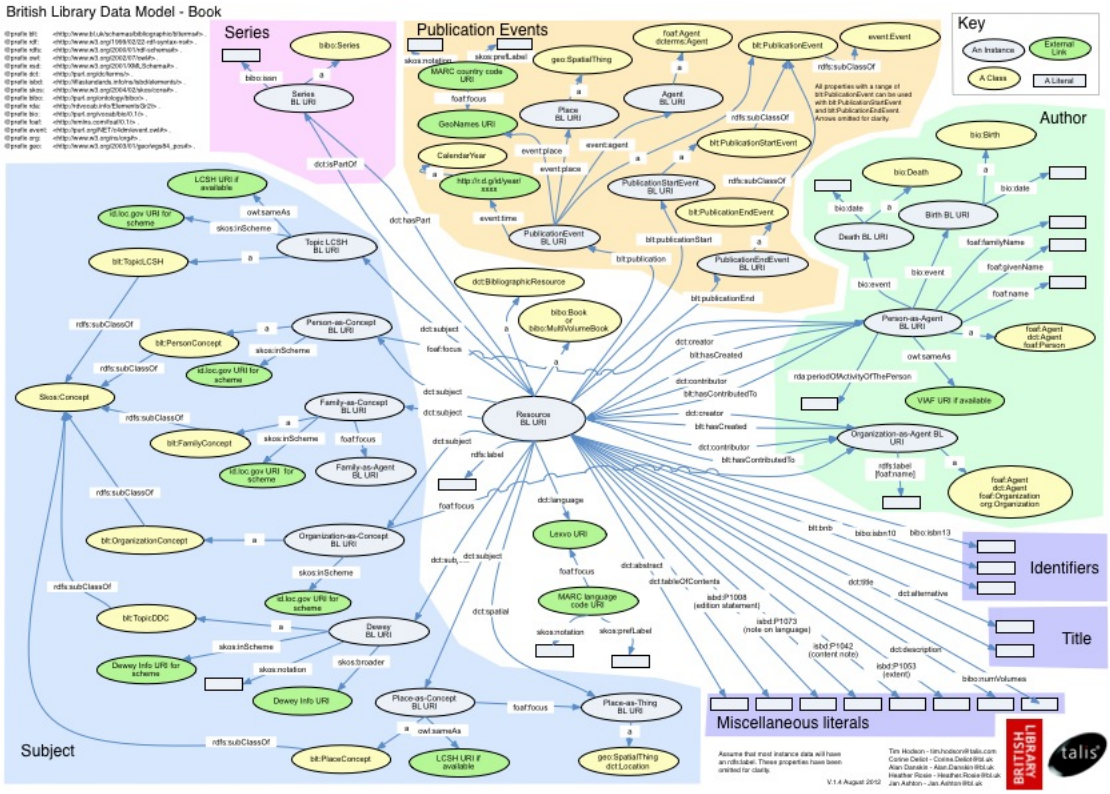
In some cases, we have chosen to use classes that appear to duplicate each other. For example, org:Organization is defined as owl:equivalentClass to foaf:Organization[8]; similarly for foaf:Agent and dcterms:Agent [9]. By using classes from different schemas our dataset can mesh with a broader range 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.[10]

We also defined our own classes and properties, documented in the British Library Terms RDF schema, [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[12] and the other for serials[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.

[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.
[11] http://www.bl.uk/schemas
[12] http://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf
[13] 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". [14] 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 LCSH, RAMEAU, etc. This approach is not without its challenges[15] 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".[16] The data is then normalized for improved matching and transformation using MARC Global, a tool developed by TMQ (The MARC of Quality)[17]. British Library-

[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/

coined URIs are generated in the MARC records and URIs linking to external datasets such as VIAF [18] and LCSH [19] 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[20]. After quality checking the RDF/XML with an open source tool, Jena Eyeball [21] , the resulting Linked Open BNB file is converted to N-Triples for loading onto our hosting platform[22]. Data dumps in both serializations (RDF/XML and N-Triples) are also loaded to our Downloads page. [23]

With respect to linking to external datasets, we chose general resources to give our data broader context, such as Geonames[24] (for country of publication), Lexvo for languages [25]. We also linked to library domain datasets: VIAF and LCSH, as already noted, but also Dewey.info[26] and the MARC Country[27] and Language[28] 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 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 [29].

The British Library selected TSO[30] 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[31] 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-sparql) is also available. The BNB is refreshed each month, both on the platform and on the British Library bulk downloads page[32].

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.

---

[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
[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

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 [33]. 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.

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[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, especially for ontologies? 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

[33] The white paper is available here: http://research.microsoft.com/apps/pubs/?id=193076
[34] http://bnb.data.bl.uk/docs

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[35] recognized the value of our approach. Our colleagues at the Danish Bibliographic Centre (DBC)[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.

[35] http://www.clir.org/pubs/reports/pub152/LDWTechDraft_ver1.0final_111230.pdf
[36] http://opensource.dbc.dk/linked-data