SEMANTIC MARC, MARC21 AND THE SEMANTIC WEB

Rob Styles
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
rob.styles@talis.com

Danny Ayers
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
danny.ayers@talis.com

Nadeem Shabir
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
nadeem.shabir@talis.com

ABSTRACT

The MARC standard for exchanging bibliographic data has been in use for several decades and is used by major libraries worldwide. This paper discusses the possibilities of representing the most prevalent form of MARC, MARC21, as RDF for the Semantic Web, and aims to understand the tradeoffs, if any, resulting from transforming the data. Critically our approach goes beyond a simple transliteration of the MARC21 record syntax to develop rich semantic descriptions of the varied things which may be described using bibliographic records. We present an algorithmic approach for consistently generating URIs from textual data, discuss the algorithmic matching of author names and suggest how RDF generated from MARC records may be linked to other data sources on the Web.

Keywords

MARC, MARC21, RDF, Semantic Web, Data Conversion, Inferred Semantics.

1. INTRODUCTION

A great deal of data exists as strings of text in structured form within binary file formats. Imagine all the ID3 tags on MP3s or all the EXIF tags in jpeg images. A more complex variation is the bibliographic data created by the hard work of generations of librarians; getting at that data is the purpose of this paper. The principles described here, though, are equally applicable to any form of data where humans are left to infer meaning from literal strings.

The MARC standard for exchanging data has been around for more than 30 years. It is a structured binary format that has allowed libraries to exchange bibliographic data very successfully. So successfully, in fact, that the Library of Congress and British Library have around 10 million records in this form each. Most national libraries have a similar number. OCLC Worldcat, a US database of libraries' information has many tens of millions. This data is not readily available for re-use outside of the library community. Talis has, for more than 40 years, maintained a database of such bibliographic records currently numbering in the tens of millions, a mixture of contributed data from libraries and commercial data from suppliers.

The Semantic Web, a web of data linked through the use of URIs and accessible over HTTP, offers the opportunity to create large, interconnected sets of data.

This paper aims to discuss the possibilities of representing the most prevalent form of MARC, MARC21, as RDF for the Semantic Web.

2. MARC21

MARC21 is used to describe several different types of record in library catalogues. Bibliographic records describe publications, Authority records list the known forms of authors' names, titles or subject headings. All of the major library management systems in use in english-speaking countries are able to import and export data in this form.

There are other flavors of MARC; DanMARC, UNIMARC and UKMARC are just some examples. The different MARC standards all share an underlying record syntax, ISO2709, but vary in the semantics assigned to different parts of the record. They differ in the level of granularity at which they store data, a single name field versus separate fore and surnames being one example, and also in where they locate data within a record - that is what meaning is assigned to each position.

China, an increasing source of online knowledge due to their massive digitization projects, use a mixture of MARC21, CMARC and CNMARC. With the volume of data available in MARC21 and the global connectivity provided by the internet, MARC21 is rapidly becoming the lingua franca for libraries globally. The techniques described in this paper are equally applicable to all flavors of MARC as well as other data formats.

```
00673nam a2200217 a 45040010033000000030009000330
05001700042008004100059015001900100020001700119035
00170013604000310015308200160018410000190020024500
62002192600033002813000020003146500060003346500031
003946550030004259cbbe7fc3a7346d99c281979d45b6
79cUK-BiTAL20050705133033.0990831s1999    enk
j      000 ||eng|d  aGB99Y57412bnb  a0747542155 :
  a()0747542155  aStDuBDScStDuBDSdUK-BiTAL04a823.
9142211 aRowling, J. K.00aHarry Potter and the pr
isoner of Azkaban /cJ.K. Rowling. aLondon :bBloom
sbury,c1999.  a317p. ;c21 cm. 0aPotter, Harry (Fi
ctitious character)vJuvenile fiction. 0aWizardsvJ
uvenile fiction. 7aChildren's stories.2lcsh
```

Figure 1. A MARC21 record in ISO2709 Syntax.

A typical MARC21 record will contain the title of the publication, the name of the creator and other contributors and often subject classifications. The underlying format was designed in the late 60's as an interchange format between machines, but due to its longevity the language of the format has become the language of the data also.

Many people find the format difficult to understand, but once the basic structure is understood it can be put to the back of the mind as there are many free and open-source tools for manipulating and getting access to this data format.

A typical MARC21 record in its native form is shown in figure 1, this form is obviously meant to be parsed by machines. A more readable representation of the same record is shown in figure 2.

```
=LDR  00673nam a2200217 a 4504
=001  9cbbe7fc3a7346d99c281979d45b679c
=003  UK-BiTAL
=005  20050705133033.0
=008  990831s1999\\\\enk  j\\\\\\000\||eng|d
=015  \\$aGB99Y5741$2bnb
=020  \\$a0747542155 :
=035  \\$a()0747542155
=040  \\$aStDuBDS$cStDuBDS$dUK-BiTAL
=082  04$a823.914$221
=100  1\$aRowling, J. K.
=245  00$aHarry Potter and the prisoner of
Azkaban /$cJ.K. Rowling.
=260  \\$aLondon :$bBloomsbury,$c1999.
=300  \\$a317p. ;$c21 cm.
=650  \0$aPotter, Harry (Fictitious character)
$vJuvenile fiction.
=650  \0$aWizards$vJuvenile fiction.
=655  \7$aChildren's stories.$2lcsh
```

Figure 2. A MARC21 record shown in human-readable form.

In this form we can see the structure, down the left a series of fields, each given a 3 character identifier. We can tell from those marked '=650' that field identifiers are not necessarily unique within a record. The later fields, from '=015' onwards have two characters of data followed by a number of subfields - each subfield starts with a $ and a single character code such as '$a'. A little knowledge of the publication in question allows us to pick out the author's name (100 $a and 245 $c), the title (245 $a), some subjects (650 $a), the publisher (260 $b) and the year of publication (260 $c). The 001 field is a special case in the standard, it is both mandatory and non-repeatable and contains a control number for the record that should be unique within a collection of records.

In addition to the field and subfield structure, the order that fields occur can also be important. The ISBN (seen here in 020 $a), for example, may be repeated; the first occurrence being the main ISBN of the material, subsequent ISBNs being related to it in some way.

This record structure is so well established within the library domain that many people speak in terms of the field and subfield codes when discussing the art of cataloguing. The specification undergoes regular revision; the MARC21 standard is the latest revision in a long line of revisions previously called USMARC. The standard is managed by the Library of Congress [8].

That MARC21 standard is solid, well used and well understood standard within the library domain.

3. RDF AND SEMANTIC WEB

The Resource Description Framework (RDF) is a suite of W3C specifications[*] which offers a model of information based on logical relationships between entities known as resources. These relationships are expressed as sets of statements (triples) in the form *subject*, *predicate*, *object*, where the object provides a value of the subject for the given property (predicate). Because the object of one statement may be the subject of another, sets of statements may also be considered graph structures, in which the subjects and predicates appear as nodes linked by (property) arcs.

RDF's notion of a resource is closely related to the resource as found on the Web identified by Uniform Resource Identifiers (URIs). A node in an RDF graph may be a URI with an optional fragment identifier (this generalization is known as a URI reference), a literal value, or blank (having no form of identification independent of the local graph). Properties are always URI references.

This core model is built upon by RDF Schema[†] (RDFS) and the Web Ontology Language[‡] (OWL), and together they can be used as a versatile yet relatively straightforward description language.

The entity-relationship model of RDF is eminently suitable for expressing data. In fact RDF stores are in many ways comparable

[*] http://www.w3.org/RDF/

[†] http://www.w3.org/TR/rdf-schema/

[‡] http://www.w3.org/TR/owl-semantics/

to relational databases, to the extent of supporting a SQL-like query language, SPARQL[*]. A key difference is that URIs provide a global naming scheme which allows immediate interoperability between any data sets expressed in RDF. Because in principle anything can be considered a resource, RDF can directly describe any thing, such as real-world objects, people, places or even abstract concepts.

The Web is primarily a massive, distributed document repository. It appears as a cohesive whole thanks to the use of URIs as identifiers, which allow links between documents to be created within a global context. The HTTP protocol allows navigation of this space, either directly in the browser or with the aid of search engines which index documents while crawling the hypertext space. Relational databases are commonly used to store information on the Web, but typically that information is only exposed as HTML, destined for human consumption through the browser.

RDF has various serialization formats (such as RDF/XML) through which it can be exposed directly on the Web as documents. However, because it uses URI references as identifiers, a greater degree of integration with the Web is possible compared to typical data representation languages. The web of documents that prevails today uses hyperlinks to create relationships between documents, the relationships between resources in RDF are exactly the same, allowing the web of documents to interlink with a new web of data or semantic web.

On the traditional Web, where a document resource is identified with a HTTP URI, the URI may be 'dereferenced' using the HTTP protocol to obtain a representation of that resource. Although there are certain complications, the same basic follow-your-nose mechanism can be used to link data. A HTTP URI for an entity (or relationship) can be dereferenced to find out more about an entity (or relationship).

The ontological constructs of RDFS and OWL also allow semantic 'linkage' to be expressed. A particularly useful example is the owl:sameAs property, which can be used to state that a resource with one URI is the exact same individual entity as that identified with another URI. With these technologies, data from different sources can easily be connected together and treated as a whole.

Clearly a major part of the success of the traditional Web is due to the benefits people gain through interlinking both within their their own documents and into the wider system. For minimal cost, their material becomes discoverable and accessible to a global audience, and the information it contains is augmented by its linkage to and from other documents. The intuition of the Semantic Web is that benefits of a similar nature (though potentially of significantly greater magnitude) will be apparent due to the network effect of linked data.

While it is still early days, the work of the Linking Open Data community [3], which has linked together 30 or so sizable data sets in this fashion (as of October 2007, comprising two billion RDF triples, interlinked by around 3 million RDF links), suggests that not only is this approach technically feasible, but that it will enable a whole new range of data-driven applications.

4. APPROACH

The work described in this paper was intended to develop an understanding of how the resources described in bibliographic data could be described using RDF and to understand what tradeoffs, if any, would result from transforming the data.

The work was carried out over a period of six months with input from a number of interested parties within the development community at Talis and brought together by Rob Styles and Nadeem Shabir.

Described here is how our understanding developed over time from a transliteration of the MARC21 record syntax to a rich semantic description of the many and varied things described by bibliographic records.

5. STRAIGHT-FORWARD REPRESENTATION

An obvious first step is to perform a simple mapping from the MARC format to RDF in much the same way as the text format shown earlier makes a simple human readable form.

Work to do this was done by Davis [5] and this provides a basic and straight-forward representation of MARC in RDF. Davis' work allowed for the record to be reconstructed accurately from the RDF, without that constraint we can produce a very simple transliterated representation of our record. An example of that is shown in figure 3 with some fields removed, the full record is shown in this form in Appendix A.

The root node of this example is the record itself, a blank node in this example. The model, being a transliteration is really no more than the original record transformed syntactically to turtle.

The primary problem with this representation is that the semantics of particular predicates depend on data close-by in the structure, for example the final field, 655, contains "Children's Stories.". That *lcsh* is the source of the term, the Library of Congress Subject Headings, is deduced from indicator 2 being "7". This kind of complex interplay with values in indicators establishing the meaning of data in the fields is common throughout the MARC standards.

This representation also suffers from quite severe readability issues for developers, relying on a knowledge of the MARC21 standard to understand the meaning of the data. Ideally we want to go beyond this, so we will look at making the properties more meaningful to humans in the next section. This is key to the semantic nature of the ontology - the simple model above

[*] http://www.w3.org/TR/rdf-sparql-query/

```
@base <http://example.com/a_marc_record> .
@prefix marc21: <http://example.com/marc21#> .
[]
  <marc21:LDR> “00673nam a2200217 a 4504”;
  <marc21:001>
“9cbbe7fc3a7346d99c281979d45b679c”;
  <marc21:005> “20050705133033.0”;
  <marc21:008> “990831s1999  enk  j    000 ||eng|
d”;
  <marc21:020> [
    <marc21:a> “0747542155 :”
  ];
  <marc21:100> [
    <marc21:ind1> “1”;
    <marc21:a> “Rowling, J. K.”
  ];
  <marc21:245> [
    <marc21:ind1> “0”;
    <marc21:ind2> “0”;
    <marc21:a> “Harry Potter and the prisoner of
Azkaban /”;
    <marc21:c> “J.K. Rowling.”
  ];
  <marc21:650> [
    <marc21:ind2> “0”;
    <marc21:a> “Potter, Harry (Fictitious
character)”;
    <marc21:v> “Juvenile fiction.”
  ], [
    <marc21:ind2> “0”;
    <marc21:a> “Wizards”;
    <marc21:v> “Juvenile fiction.”
  ];
  <marc21:655> [
    <marc21:ind2> “7”;
    <marc21:a> “Children's stories.”;
    <marc21:2> “lcsh”
  ] .
```

Figure 3 - Straight-Forward Representation

provides no semantics over and above the original MARC21 format and, as such, has little value.

The data also contains display punctuation. This is a historic artifact of MARC21 called ISBD punctuation, allowing the data to be concatenated for printing without programmatic formatting. It is widely accepted that this is not ideal. Removing the punctuation requires knowledge of which fields should have specific characters removed from front and back. The Library of Congress has rules for this contained in their MARC to MODS[*] stylesheets.

6. READABLE REPRESENTATION

The representation above is a very basic transliteration of the MARC standard. The alpha-numeric codes for fields and subfields still dominating. However, the Library Of Congress provide human-readable labels for the field and subfield codes and these are used in some tools to provide a more meaningful representation.

One of the benefits of RDF is the ability to bring data together more effectively than with other formats. Sticking with the alphanumeric codes alone, all types of MARC would be represented by very similar predicates, differing only in the base URI. While this makes no difference to the machine interpretation of the data it provides difficult and confusing semantics for developers working with the data. By mapping the codes to

```
@base <http://example.com/a_marc_record> .
@prefix marc21: <http://example.com/marc21#> .
[]
  <marc21:controlNumber>
“9cbbe7fc3a7346d99c281979d45b679c”;

#Following data comes from fixed positions in the
Leader
  <marc21:recordStatus> “New”;
  <marc21:recordType> “Language material”;
  <marc21:bibliographicLevel> “Monograph/item”;
  <marc21:encodingLevel> “Full”;

#Following data comes from fixed positions in 008
  <marc21:recordCreated>
“1999-08-31”^^xsd:dateTime;
  <marc21:publicationStatus> “Published”;
  <marc21:placeOfPublication> “England”;
  <marc21:language> “English”;
  <marc21:targetAudience> “Juvenile”;
  <marc21:festschrift> “No”;

#Following data comes from other control fields
  <marc21:controlNumberIdentifier> “UK-BiTAL”;
  <marc21:recordUpdated>
“2005-07-05T13:30:33Z”^^xsd:dateTime;
  <marc21:nationalBibliographyNumber> [
  <marc21:number> “GB99Y5741”;
  <marc21:sourceOfNumber> “bnb”;
  ];
  <marc21:isbn> “0747542155”;
  <marc21:deweyDecimalClassification> “823.914”
  <marc21:associatedPersonalName> “Rowling, J.
K.”;
  <marc21:title> “Harry Potter and the prisoner
of Azkaban”;
  <marc21:statementOfResponsibility> “J.K.
Rowling.”;
  <marc21:placeOfPublication> “London”;
  <marc21:dateOfPublication>
“1999”^^xsd:dateTime;
  <marc21:publisher> “Bloomsbury”;
  <marc21:physicalExtent> “317p.”;
  <marc21:physicalDimensions> “21 cm”;
  <marc21:topicalTerm> [
  <marc21:sourceOfTerm> “LCSH”;
  <marc21:term> “Potter, Harry (Fictitious
character)”;
  <marc21:formSubdivision> “Juvenile fiction.”;
  ], [
  <marc21:sourceOfTerm> “LCSH”;
  <marc21:term> “Wizards”;
  <marc21:formSubdivision> “Juvenile fiction.”
  ];
  <marc21:genre> [
  <marc21:sourceOfTerm> “LCSH”;
  <marc21:term> “Children's stories.”;
  ] .
```

Figure 4 - Readable Representation

[*] http://www.loc.gov/standards/mods/mods-userguide-intro.html

bibliographic terms, and other MARC flavors doing the same, it is possible to create a more interoperable ontology - this is what was hoped would happen with MODS [10] and is also the hope for the combined work of DCMI and RDA working groups [4].

Without taking other, existing, schemas into account (at this point) and working with the Library of Congress descriptions we can provide a more semantically rich representation of the data. This involves interpreting the data in certain parts as well as the structure of the record and also requires some fixed format data, such as the Leader and 008 to be parsed.

Doing this, for an incomplete example set from the record above, we arrive at a simpler and more readable record as shown in figure 4. The full record is given in Appendix B.

This more readable representation uses terms we can interpret, apart from 'festschrift' perhaps which may require a lookup[*]. More importantly, though, the data tied up in fixed positions of fields has been drawn out and the areas where the meaning of a predicate depended on data close-by the predicates have been disambiguated.

An ontology based on the full Library of Congress MARC21 specification would be valuable in it's own right, and MODS is very close being just that, as is the full dublin-core specification. However, the data still lies in literal values, meaning that each record remains an island of data. The same concepts, such as *Harry Potter*, *London* or *J. K. Rowling* occurring in many records will not be related.

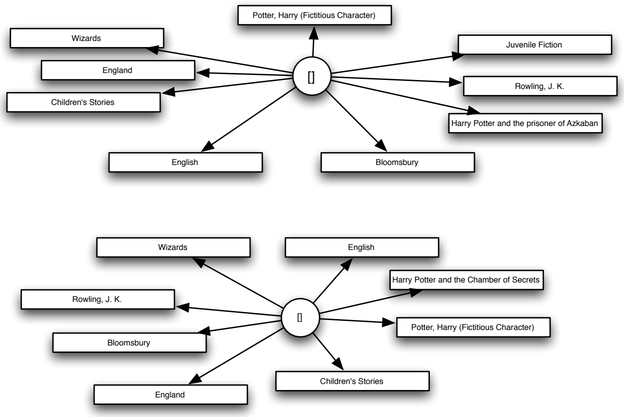
Potter, Harry (Fictitious Character)
Wizards
Juvenile Fiction
England
Rowling, J. K.
Children's Stories
Harry Potter and the prisoner of Azkaban
English
Bloomsbury
Wizards
English
Harry Potter and the Chamber of Secrets
Rowling, J. K.
Potter, Harry (Fictitious Character)
Bloomsbury
Children's Stories
England

Figure 5 - Visualization of two translated MARC21 records.

A visualization of two translated records, figure 5, shows the problem clearly. What we can see is that each record forms its own island of data. Concepts which we would expect to be shared between records are not shared, leaving little or no opportunity to navigate the data through simple links.

7. LINKS NOT LITERALS

The readable representation provides many benefits, but still fails to exploit the full potential of the data. The use of literals, while it is syntactically RDF, doesn't give identity to the concepts and doesn't allow them to be referenced. That is, as literals, I can't refer to them in other triples or add additional information about them.

To rectify this we need to replace literal values with URIs representing the concepts that the literals provide names for. This means creating a URI to refer to the resource that we, as people, infer from the literal string. We then use the literal string as a label for the URI.

This is good confirmation of the principles behind the work done at DERI on MarcOnt [12], a similar ontology for MARC21. MarcOnt specifies that subjects should be SKOS concepts and the people involved with a resource should be represented as FOAF person. Other things, such as title, remain string literals.

8. SEMANTIC REPRESENTATION

The URIs used to represent the resources could be arbitrarily assigned using GUIDs or some other mechanism. However, to match the same URI for the same resource would then require a query. As we are processing large volumes of data using a parallel processing architecture we decided to create URIs algorithmically.

If we had assigned URIs rather than computing them then as we parse the data and come across the title *Stardust* we would have to look up the identifier for it, and if there was not one create a new one and register it. This would require a huge number of lookups, a centralized store and distributed locking mechanisms for updates to the store.

Not only did we want to be able to process large volumes of this data in parallel and arrive at the same URIs even when processing data in distributed locations around the net, we also wanted to allow others to process their own data and arrive at a set of URIs that could link in without them having a dependance on us.

To allow for this we have started with a simple hashing algorithm to arrive at a key from a given piece of text. The intention of developing a hash is that different strings that represent the same resource will converge on the same hashed value, while strings representing different resources will diverge on different hashed values.

The algorithm is currently the same for names, titles, subjects and other types of text. We feel confident that there are opportunities to develop optimized algorithms for names and subject terms. This could allow a higher degree of accuracy in matching texts that represent the same concept.

Our initial algorithm was formulated on data mostly in English with a few entries in other Western European languages. Different techniques would be required for different language groups.

Our starting point is to remove all punctuation and whitespace and lower-case the string. When appended to a base URI for each type of datum this gives a simple URI that will be consistent for the main variations that occur in bibliographic data.

[*] http://en.wikipedia.org/wiki/Festschrift

Taking the author's personal name from our previous example - *Rowling, J. K.* we get a hashed value *rowlingjk* which would also be generated from other common forms of the name such as *Rowling J K* or *Rowling, J.K.*. We will discuss possibilities for dealing with more substantial variations such as *Joanne K. Rowling* in section 10 on authority data.

The hash is then appended to a base URI for the type of resource, so for J. K. Rowling we create *http://example.com/people/rowlingjk#self*. The title of Rowling's book, *Harry Potter and the Chamber of Secrets* creates *http://se.../titles/harrypotterandthechamberofsecrets#self*.

Doing this for names, places, organizations, titles and other literals for two example records we arrive at the graph shown in figure 6. Instead of the islands of data that we saw previously in figure 5 the two records have become entwined, sharing many concepts.

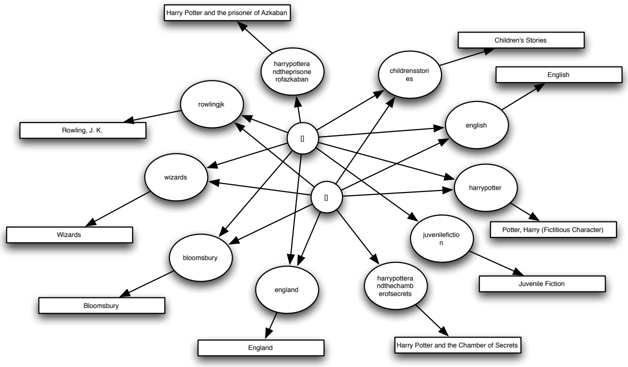
Harry Potter and the prisoner of Azkaban
harrypotterandtheprisonerofazkaban
Children's Stories
childrensstories
English
rowlingjk
Rowling, J. K.
english
[]
wizards
[]
harrypotter
Wizards
Potter, Harry (Fictitious Character)
juvenilefiction
bloomsbury
Juvenile Fiction
england
harrypotterandthechamberofsecrets
Bloomsbury
England
Harry Potter and the Chamber of Secrets

Figure 6 - Visualization of two interpreted MARC21 records.

The current algorithm does not address ordering differences, but initial work has been done on a hash that re-orders the characters from *rowlingjk* to *gijklnorw*. This work shows that this hash also has a very high degree of uniqueness and addresses many ordering issues. We have decided against using it currently as we would rather have a higher number of unmatched authors than a higher number of incorrectly matched authors.

9. LINKING TO OTHER SOURCES

The resulting graph of resources now lends itself readily to linking into other data sources. The Library of Congress, for example, is working towards releasing their Subject Headings on the web using SKOS. If this were done using the subject terms and a known base URI it would be trivial to generate those links during the conversion from MARC21 to RDF without having to perform a lookup.

The resources also represent people and organizations, so a link in with the FOAF schema and cross-referencing with DBPedia [1] would be very beneficial. The place names could also be linked in with Geonames[*].

The RDF Book Mashup [2] maintains a simpler schema, using dublin-core, so compatibility could be achieved using a fairly simple transform.

10. ADDING AUTHORITY DATA

With any algorithmic approach there will be flaws that need to be corrected through human intervention. In the case of MARC data it is to be expected that authors' names, titles and other textual data will be represented slightly differently between different editions of a work. This will result in the same resource being represented by more than one URI.

In the library domain there has long been a solution for this problem in Authority records. Authority records document the variant forms of authors' names, titles and subject headings and relate them to a single authorized form. With a little parsing these entries can quickly become cross-references or "see" relationships in the RDF, unifying the multiple resource URIs.

When data is generated in several different places from data with different levels of completeness the "see" relationships allow that data to be combined with other data more completely. In a single graph they could be removed and all occurrences replaced with a single, authoritative URI. With distributed data sets coming together later the "see" relationship maintains its value.

This is a very different approach to that taken by Hickey et al [7] where authority data was used to clean data before processing it.

The authority data may also contain additional information about relationships between authors' names; one example being that Iain Banks also publishes under the name Iain M Banks, another common example being that Mark Twain was the pen name of Samuel Clemens. These relationships are between different resources rather than different URIs representing the same resource, so require a "see also" relationship rather than a "same as" relationship.

Authority data is available for variant forms of titles and equivalent or related subject terms as well as names of people and organizations.

Going back to our first example record, we find J. K. Rowling referred to as both "Rowling, J. K." and "J.K. Rowling". Another common form is "Rowling, Joanne K.". URIs can be generated from all of these, cross-referencing the different forms. This is what name authorities have been doing for many decades, but in a set of records separate from the data about items.

```
=100  1/$a Rowling, J. K.
=400  1/$a Rowling, Joanne K.$q (Joanne Kathleen)
=400  1/$a Rowling, Jo
=400  1/$a Scamander, Newt
=400  1/$a Whisp, Kennilworthy
=400  1/$a Roling, G’e. K e
=400  1/$a Rowlingová, Joanne K.
=400  1/$a Roling, Dz´h`. K.
```

Figure 7 - Extract from a MARC21 Authority Record

[*] http://www.geonames.org/

Figure 7 shows an extract of a name authority record for J. K. Rowling showing a number of variant forms as well as “Scamander, Newt” and “Whisp, Kennilworthy”, two pseudonyms under which Rowling published companion works to the Harry Potter series.

The extract shown in Figure 7 can be used to add to the graph shown in Figure 6. The resulting graph having an additional 7 nodes allowing discovery of J. K. Rowling through alternative forms of her name. This is shown in figure 8.

11. ADDING LESS PRECISE URIS

Generating URIs algorithmically from data works well as long as everyone has the same amount of data to work with. If we take authors’ names as our example, it is common in MARC data to disambiguate authors with the same name using the year of birth.

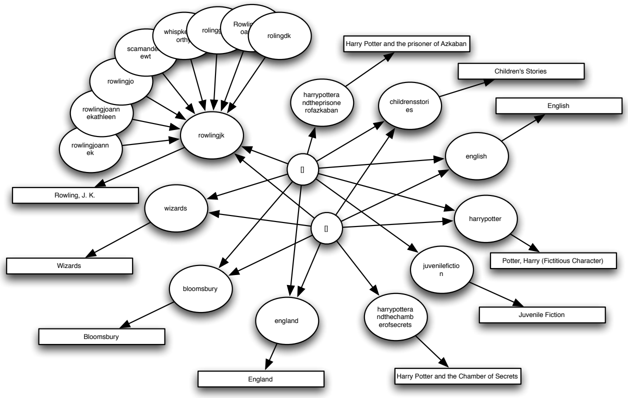

Figure 8 - Graph with Authority Data added.

If the year of birth is not available in the data then it is not possible to generate the same URI as that generated from data where the year is available.

Taking the example of Samuel Roberts, there are a number of notable authors of that name including a writer on psychoanalysis born in 1961 and a writer on classical music born in 1962.

It is common for either of them to be referenced simply as “Roberts, Samuel” resulting in the same URI.

A URI generated from “Roberts, Samuel” intends to refer to “Roberts, Samuel, 1961”, “Roberts, Samuel, 1962” or one of the many other authors of that name, but will not match any unambiguously.

It is not possible to generate the more precise reference from the less precise one; however it is possible to generate the less precise reference when we have the more precise form in hand.

When we encounter either “Roberts, Samuel, 1961” or “Roberts, Samuel, 1962” we can create a relationship with the less precise “Roberts, Samuel” providing the opportunity to disambiguate references to the less precise form later, either through algorithms similar to Soler’s [11] or through manual intervention.

12. IDENTIFYING “COULD BE” MATCHING NAMES ALGORITHMICALLY

A further avenue for exploration would be to explore the distance between the hashed values for various names. String similarity metrics such as Jaro Winkler or Levenshtein [citation required] could provide indications of typographical errors in the source data.

The combination of an author’s name and the title of their work or works provides a highly distinctive signature that could be used either to select candidates for comparison or to provide a further validation of close matches found in the corpus as a whole.

13. FRBR

FRBR [13], Functional Requirements for Bibliographic Records, is a conceptual data model that, amongst other things, provides four concepts around which attributes can be grouped. The four levels are best described through an example as their definition is not entirely concrete.

Take Bach’s Partita No.1 in B Minor, say you have some commercially printed sheet music for it in your hand. What you have in your hand is, in FRBR terms, an Item so it has lots of attributes specific to that one thing you hold in your hand - like the coffee stain your sister made on it. There are also lots of other, near identical items that the printer made at the same time from the same pattern, this pattern is the Manifestation in FRBR. The manifestation has other attributes like the ISMN*, the date it was published and so on. Then there are many different printed versions which, although having different ISMNs, being printed by different printers and at different times, contain the exact same musical notes. That set of notes put down in that way is known as the Expression. Finally, going all the way back to around 1728 we have the idea in Bach’s mind which we call the Work.

This model requires some thought and, despite being published in 1998 has only recently started to gain traction. The boundaries that define when two things constitute different works, expressions and manifestations are not well understood and there is still work to do to test the model [9]. Figure 9 shows an informal representation of how the model might apply to Bach’s Partita No. 1.

Despite needing more work, FRBR already has great value in its current form in relating things that would otherwise remain separate.

* http://en.wikipedia.org/wiki/ISMN

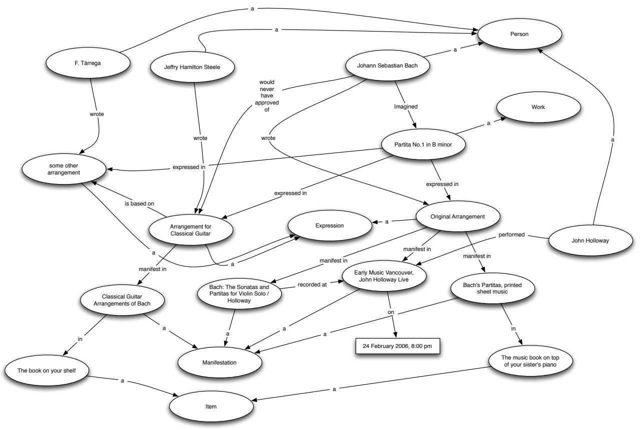

Figure 9 - Explanatory Diagram of FRBR

The FRBR model was described in RDF by Davis and Newman[*]. Work by the Library of Congress [6] has been used, along with Davis' FRBR ontology, to establish FRBR resources relationships in the data.

The Library of Congress work provides matching, sorting and display criteria for grouping MARC21 records according to the Work, Expression, Manifestation, Item model. We took the matching criteria and used those to generate URIs from each record to represent Work, Manifestation and Expression.

The Library of Congress rules for matching and author, for example, say to take the following fields and subfields and comparing them, 100$a$b$c$d (or) 110$a$b$c$d (or) 111$a$c$d $n$q and compare them paying no attention to case, whitespace, punctuation, brackets or parentheses. Instead of matching on this we use this to generate a URI that represents the author, as previously described. The same process is used in generation of a URI to represent the title.

A Work is defined by the Library of Congress as anything that matches both the author and the title. We can apply the same logic to generate a URI to represent the Work by combining our author and title hashes. Earlier we created an author URI for J. K. Rowling of

.../people/rowlingjk#self

and a title URI for *Harry Potter and The Chamber of Secrets* of

.../titles/harrypotterandthechamberofsecrets#self.

By simply combining the two hashes we create a URI for the work of

.../works/rowlingjkharrypotterandthechamberofsecrets#self

As we move down into Expression and Manifestation URIs the URIs get extremely long. To get back to a more usable length we simply applied the MD5 hashing algorithm to our longer types of URI giving a shorter URI like this

.../works/7317d9412ec8b804e00bfe9989d10521#self

This URI has all the convenience of a GUID style generated URI but can be generated by different groups in disconnected environments with a high degree of success in generating the same URI even from data with variations.

As different organizations use their own interpretations of the FRBR entities it would be possible for each to use it's own predicates for their interpretation of FRBR. This would allow many interpretations to peacefully co-exist within the same set of data. Anyone wishing to treat all interpretations equally would be able to associate the different predicates using owl:sameAs.

A full listing of the model generated from our two example records is given in Appendix C.

14. DEALING WITH CORRECTIONS AND PERSISTENT URIS

Once the data is published it is inevitable it will require corrections. Leaving corrections to the literal values aside, corrections of URIs fall into two classes: where two different URIs have been assigned to represent what is actually a single resource; and where the same URI has been used to describe what is really two distinct resources.

Disambiguation is required when a single URI has been used to reference two concepts that it later transpires are distinct. This has strong similarities to the discussion in section ?? above on adding less precise URIs. In that section we were deliberately adding ambiguous URIs into the data to facilitate greater discoverability.

We need to use the same disambiguation technique when we discover that the same URI has been used for two discreet concepts.

This is directly analogous to the way in which wikipedia has evolved disambiguation pages[†] where a term is equally applicable to more than one topic.

Convergence deals with the opposite of disambiguation - situations where more than one URI has been generated for what is later decided is a single resource.

Convergences are expected to occur through three main mechanisms; refinements of the algorithms, addition of data from other sources and human intervention in the data.

Where convergent URIs are discovered we intend to migrate relationships to a single URI and maintain redirects for any URIs that have been superseded.

15. OTHER ONTOLOGIES

The work described in this paper focussed on understanding what was possible with MARC21 data rather than broader interoperability issues. Next stages under consideration by the

[*] http://vocab.org/frbr/core

[†] http://en.wikipedia.org/wiki/Wikipedia:Disambiguation

team are to compare and contrast the ontology resulting from this work with other efforts such as MarcOnt, Bibliontology [citation required] and the RDF Book Mashup.

The team maintain a watching brief on the work being done by DCMI and RDA working groups and the work being done by the Library of Congress and other organizations such as the Internet Archive's Open Library project.

16. SUMMARY

The work done here provides the start of an approach that would allow decades of interesting and culturally valuable data to be transformed into a part of the web of data.

Further research is needed to arrive at hashing algorithms that maximize the matching of resources from literals strings and the teams is interested in hearing suggestions on that.

It is the team's hope to release a substantial set of data onto the web using these approaches during 2008 at which time the team will be looking to link both to other data sets and other ontologies.

17. REFERENCES

[1] Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R. and Ives, Z. (2007) DBpedia: A Nucleus for a Web of Open Data. In 6th International Semantic Web Conference, Busan, Korea.

[2] Bizer C., Cyganiak R., Gau§ T. (007) The RDF Book Mashup: From Web APIs to a Web of Data http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/Bizer-ESWC2007-RDFbookmashup.pdf

[3] Bizer C., Heath T., Ayers D., Raimond Y. (2007) Interlinking Open Data on the Web. In Demonstrations Track, 4th European Semantic Web Conference (ESWC2007), Innsbruck, Austria. http://www.eswc2007.org/pdf/demo-pdf/LinkingOpenData.pdf

[4] Clayphan R. et al (2007) Data Model Meeting http://www.bl.uk/services/bibliographic/meeting.html

[5] Davis I. (2005) MARC Transliteration http://iandavis.com/blog/2005/12/marc-transliteration

[6] Delsey T. (2002) Functional Analysis of the MARC 21 Bibliographic and Holdings Formats http://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html

[7] Hickey T., O'Neill E., Toves J. (2002) Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR) http://www.dlib.org/dlib/september02/hickey/09hickey.html

[8] Furrie, B. (2003) Understanding MARC Bibliographic: Machine-Readable Cataloging http://www.loc.gov/marc/umb/

[9] Madison O. et al (2008) On The Record: Report of The Library of Congress Working Group on the Future of Bibliographic Control http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf

[10] Moen W. (2005) Metadata Interaction, Integration, and Interoperability http://www.unt.edu/wmoen/presentations/MD_Interoperability_wem_June2005.ppt

[11] Soler J. (2007) Separating the articles of authors with the same name DOI=10.1007/s11192-007-1730-z

[12] Synak M., Kruk S. R. (2005) MarcOnt Initiative - the Ontology for the Librarian World http://www.marcont.org/marcont/pdf/ms_eswc2005marcont.pdf

[13] Yee M., (2005) FRBRization: a Method for Turning Online Public Finding Lists into Online Public Catalogs http://repositories.cdlib.org/postprints/715

APPENDIX A - STRAIGHT-FORWARD REPRESENTATION

```
@base <http://example.com/a_marc_record> .
@prefix marc21: <http://example.com/marc21#> .
[]
  <marc21:LDR> “00673nam a2200217 a 4504”;
  <marc21:001> “9cbbe7fc3a7346d99c281979d45b679c”;
  <marc21:003> “UK-BiTAL”;
  <marc21:005> “20050705133033.0”;
  <marc21:008> “990831s1999  enk  j    000 ||eng|d”;
  <marc21:015> [
    <marc21:a> “GB99Y5741”;
    <marc21:2> “bnb”
  ];
  <marc21:020> [
    <marc21:a> “0747542155 :”
  ];
  <marc21:035> [
    <marc21:a> “()0747542155”
  ];
  <marc21:040> [
    <marc21:a> “StDuBDS”;
    <marc21:c> “StDuBDS”;
    <marc21:d> “UK-BiTAL”
  ];
  <marc21:082> [
    <marc21:ind1> “0”;
    <marc21:ind2> “4”;
    <marc21:a> “823.914$221”
  ];
  <marc21:100> [
    <marc21:ind1> “1”;
    <marc21:a> “Rowling, J. K.”
  ];
  <marc21:245> [
    <marc21:ind1> “0”;
    <marc21:ind2> “0”;
    <marc21:a> “Harry Potter and the prisoner of Azkaban /”;
    <marc21:c> “J.K. Rowling.”
  ];
  <marc21:260> [
    <marc21:a> “London :”;
    <marc21:b> “Bloomsbury,”;
    <marc21:c> “1999.”
  ];
  <marc21:300> [
    <marc21:a> “317p. ;”;
    <marc21:c> “21 cm.”
  ];
  <marc21:650> [
    <marc21:ind2> “0”;
    <marc21:a> “Potter, Harry (Fictitious character)”;
    <marc21:v> “Juvenile fiction.”
  ], [
    <marc21:ind2> “0”;
    <marc21:a> “Wizards”;
    <marc21:v> “Juvenile fiction.”
  ];
  <marc21:655> [
    <marc21:ind2> “7”;
    <marc21:a> “Children's stories.”;
    <marc21:2> “lcsh”
  ] .
```

APPENDIX B - READABLE REPRESENTATION

```
@base <http://example.com/a_marc_record> .
@prefix marc21: <http://example.com/marc21#> .
[]
  <marc21:controlNumber> “9cbbe7fc3a7346d99c281979d45b679c”;

  #Following data comes from fixed positions in the Leader
  <marc21:recordStatus> “New”;
  <marc21:recordType> “Language material”;
  <marc21:bibliographicLevel> “Monograph/item”;
  <marc21:encodingLevel> “Full”;

  #Following data comes from fixed positions in 008
  <marc21:recordCreated> “1999-08-31”^^xsd:dateTime;
  <marc21:publicationStatus> “Published”;
  <marc21:placeOfPublication> “England”;
  <marc21:language> “English”;
  <marc21:targetAudience> “Juvenile”;
  <marc21:festschrift> “No”;

  #Following data comes from other control fields
  <marc21:controlNumberIdentifier> “UK-BiTAL”;
  <marc21:recordUpdated> “2005-07-05T13:30:33Z”^^xsd:dateTime;
  <marc21:nationalBibliographyNumber> [
    <marc21:number> “GB99Y5741”;
    <marc21:sourceOfNumber> “bnb”;
  ];
  <marc21:isbn> “0747542155”;
  <marc21:deweyDecimalClassification> “823.914”
  <marc21:associatedPersonalName> “Rowling, J. K.”;
  <marc21:title> “Harry Potter and the prisoner of Azkaban”;
  <marc21:statementOfResponsibility> “J.K. Rowling.”;
  <marc21:placeOfPublication> “London”;
  <marc21:dateOfPublication> “1999”^^xsd:dateTime;
  <marc21:publisher> “Bloomsbury”;
  <marc21:physicalExtent> “317p.”;
  <marc21:physicalDimensions> “21 cm”;
  <marc21:topicalTerm> [
    <marc21:sourceOfTerm> “LCSH”;
    <marc21:term> “Potter, Harry (Fictitious character)”;
    <marc21:formSubdivision> “Juvenile fiction.”;
  ], [
    <marc21:sourceOfTerm> “LCSH”;
    <marc21:term> “Wizards”;
    <marc21:formSubdivision> “Juvenile fiction.”
  ];
  <marc21:genre> [
    <marc21:sourceOfTerm> “LCSH”;
    <marc21:term> “Children's stories.”;
  ] .
```

APPENDIX C - SEMANTIC REPRESENTATION

```
@base <http://example.com/potter.rdf> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix j.0: <http://example.com/schema/bib#> .
@prefix j.1: <http://purl.org/vocab/frbr/core#> .

<http://example.com/resources/people/rowlingjk#self>
  j.0:writesAbout <http://example.com/resources/genres/childrensstories#self>, <http://example.com/
resources/topics/wizards#self>, <http://example.com/resources/topics/
potterharryfictitiouscharacter#self> ;
  j.0:publisher <http://example.com/resources/organizations/bloomsbury#self> ;
  j.1:creatorOf <http://example.com/resources/manifestations/988e45a216923b3d24e4a4a711a664#self>,
<http://example.com/resources/expressions/7a67d02fe5b1f4fccc78eb91135a7d0#self>, <http://example.com/
resources/works/d415d3e7bb88725134eb21d11718bdaa#self> ;
  j.0:seenAs "Rowling, J. K." ;
  a j.0:Person ;
  j.1:creatorOf <http://example.com/resources/manifestations/62c544b579c57dd1c1e4092d0d02a1#self>,
<http://example.com/resources/expressions/7df5817e8c75b34766169d8ade554bfe#self>, <http://example.com/
resources/works/7317d9412ec8b84e0bfe9989d1521#self> .

<http://example.com/resources/organizations/bloomsbury#self>
  j.0:publisherOf <http://example.com/resources/manifestations/988e45a216923b3d24e4a4a711a664#self>,
<http://example.com/resources/people/rowlingjk#self> ;
  j.0:seenAs "Bloomsbury," ;
  a j.0:Publisher ;
  j.0:publisherOf <http://example.com/resources/manifestations/62c544b579c57dd1c1e4092d0d02a1#self> .

<http://example.com/resources/expressions/7a67d02fe5b1f4fccc78eb91135a7d0#self>
  j.0:format <http://example.com/resources/languages/eng#self>, <http://example.com/resources/formats/
a#self> ;
  j.0:isbn <http://example.com/resources/isbns/9780747542155#self> ;
  j.0:name <http://example.com/resources/titles/harrypotterandtheprisonerofazkaban#self> ;
  j.1:creator <http://example.com/resources/people/rowlingjk#self> ;
  j.1:embodiment <http://example.com/resources/manifestations/988e45a216923b3d24e4a4a711a664#self> ;
  j.1:realizationOf <http://example.com/resources/works/d415d3e7bb88725134eb21d11718bdaa#self> ;
  a j.1:Expression .

<http://example.com/resources/titles/harrypotterandtheprisonerofazkaban#self>
  j.0:nameOf <http://example.com/resources/manifestations/988e45a216923b3d24e4a4a711a664#self>, <http://
example.com/resources/expressions/7a67d02fe5b1f4fccc78eb91135a7d0#self>, <http://example.com/resources/
works/d415d3e7bb88725134eb21d11718bdaa#self> ;
  j.0:seenAs "Harry Potter and the prisoner of Azkaban /" ;
  a j.0:Title .

<http://example.com/resources/dates/1999#self>
  a j.0:Date .

<http://example.com/resources/languages/eng#self>
  a j.0:Language .

<http://example.com/resources/works/d415d3e7bb88725134eb21d11718bdaa#self>
  j.1:subject <http://example.com/resources/genres/childrensstories#self>, <http://example.com/resources/
topics/wizards#self>, <http://example.com/resources/topics/potterharryfictitiouscharacter#self> ;
  j.0:isbn <http://example.com/resources/isbns/9780747542155#self> ;
  j.0:name <http://example.com/resources/titles/harrypotterandtheprisonerofazkaban#self> ;
  j.1:creator <http://example.com/resources/people/rowlingjk#self> ;
  <http://example.com/resources/manifestations/988e45a216923b3d24e4a4a711a664#self> <http://example.com/
resources/expressions/7a67d02fe5b1f4fccc78eb91135a7d0#self> ;
  a j.1:Work .

<http://example.com/resources/genres/childrensstories#self>
  j.0:seenAs "Children's stories." ;
  a j.0:Genre .

<http://example.com/resources/manifestations/988e45a216923b3d24e4a4a711a664#self>
  j.0:format <http://example.com/resources/languages/eng#self>, <http://example.com/resources/formats/
a#self> ;
  j.0:datePublished <http://example.com/resources/dates/1999#self> ;
  j.0:publisher <http://example.com/resources/organizations/bloomsbury#self> ;
  j.0:isbn <http://example.com/resources/isbns/9780747542155#self> ;
  j.0:name <http://example.com/resources/titles/harrypotterandtheprisonerofazkaban#self> ;
  j.1:creator <http://example.com/resources/people/rowlingjk#self> ;
  j.1:embodimentOf <http://example.com/resources/expressions/7a67d02fe5b1f4fccc78eb91135a7d0#self> ;
  a j.1:Manifestation .

<http://example.com/resources/formats/a#self>
  a j.0:Format .

<http://example.com/resources/topics/potterharryfictitiouscharacter#self>
  j.0:seenAs "Potter, Harry (Fictitious character)" ;
  a j.0:Topic .

<http://example.com/resources/isbns/0747542155#self>
  j.0:seeAlso <http://example.com/resources/isbns/9780747542155#self> ;
  j.0:seenAs "0747542155" ;
  a j.0:ISBN .

<http://example.com/resources/isbns/9780747542155#self>
  <http://example.com/resources/manifestations/988e45a216923b3d24e4a4a711a664#self> <http://example.com/
resources/manifestations/988e45a216923b3d24e4a4a711a664#self>, <http://example.com/resources/expressions/
7a67d02fe5b1f4fccc78eb91135a7d0#self>, <http://example.com/resources/works/
d415d3e7bb88725134eb21d11718bdaa#self> ;
  j.0:seeAlso <http://example.com/resources/isbns/0747542155#self> ;
  j.0:seenAs "9780747542155" ;
  a j.0:ISBN .

<http://example.com/resources/topics/wizards#self>
  j.0:seenAs "Wizards" ;
  a j.0:Topic .

<http://example.com/resources/expressions/7df5817e8c75b34766169d8ade554bfe#self>
  j.0:format <http://example.com/resources/languages/eng#self>, <http://example.com/resources/formats/
a#self> ;
  j.0:isbn <http://example.com/resources/isbns/9780747538493#self> ;
  j.0:name <http://example.com/resources/titles/harrypotterandthechamberofsecrets#self> ;
  j.1:creator <http://example.com/resources/people/rowlingjk#self> ;
  j.1:embodiment <http://example.com/resources/manifestations/62c544b579c57dd1c1e4092d0d02a1#self> ;
  j.1:realizationOf <http://example.com/resources/works/7317d9412ec8b84e0bfe9989d1521#self> ;
  a j.1:Expression .

<http://example.com/resources/isbns/0747538492#self>
  j.0:seeAlso <http://example.com/resources/isbns/9780747538493#self> ;
  j.0:seenAs "0747538492" ;
  a j.0:ISBN .

<http://example.com/resources/works/7317d9412ec8b84e0bfe9989d1521#self>
  j.1:subject <http://example.com/resources/genres/childrensstories#self>, <http://example.com/resources/
topics/wizards#self> ;
  j.0:isbn <http://example.com/resources/isbns/9780747538493#self> ;
  j.0:name <http://example.com/resources/titles/harrypotterandthechamberofsecrets#self> ;
  j.1:creator <http://example.com/resources/people/rowlingjk#self> ;
  <http://example.com/resources/manifestations/988e45a216923b3d24e4a4a711a664#self> <http://example.com/
resources/expressions/7df5817e8c75b34766169d8ade554bfe#self> ;
  a j.1:Work .

<http://example.com/resources/isbns/9780747538493#self>
  <http://example.com/resources/manifestations/988e45a216923b3d24e4a4a711a664#self> <http://example.com/
resources/manifestations/62c544b579c57dd1c1e4092d0d02a1#self>, <http://example.com/resources/expressions/
7df5817e8c75b34766169d8ade554bfe#self>, <http://example.com/resources/works/
7317d9412ec8b84e0bfe9989d1521#self> ;
  j.0:seeAlso <http://example.com/resources/isbns/0747538492#self> ;
  j.0:seenAs "9780747538493" ;
  a j.0:ISBN .

<http://example.com/resources/manifestations/62c544b579c57dd1c1e4092d0d02a1#self>
  j.0:format <http://example.com/resources/languages/eng#self>, <http://example.com/resources/formats/
a#self> ;
  j.0:datePublished <http://example.com/resources/dates/1998#self> ;
  j.0:publisher <http://example.com/resources/organizations/bloomsbury#self> ;
  j.0:isbn <http://example.com/resources/isbns/9780747538493#self> ;
  j.0:name <http://example.com/resources/titles/harrypotterandthechamberofsecrets#self> ;
  j.1:creator <http://example.com/resources/people/rowlingjk#self> ;
  j.1:embodimentOf <http://example.com/resources/expressions/7df5817e8c75b34766169d8ade554bfe#self> ;
  a j.1:Manifestation .

<http://example.com/resources/titles/harrypotterandthechamberofsecrets#self>
  j.0:nameOf <http://example.com/resources/manifestations/62c544b579c57dd1c1e4092d0d02a1#self>, <http://
example.com/resources/expressions/7df5817e8c75b34766169d8ade554bfe#self>, <http://example.com/resources/
works/7317d9412ec8b84e0bfe9989d1521#self> ;
  j.0:seenAs "Harry Potter and the chamber of secrets /" ;
  a j.0:Title .

<http://example.com/resources/dates/1998#self>
  a j.0:Date .
```