Shrinking the silo boundary: data and schema in the Semantic Web

Gordon Dunsire

16. seminar Arhivi, Knjižnice, Muzeji

28-30 Nov 2012, Poreč, Croatia

Published by Gordon Dunsire

Edinburgh 2013

Shrinking the silo boundary: data and schema in the Semantic Web

Gordon Dunsire

Abstract

Until recently, the methodology of publishing legacy bibliographic metadata as linked open data has been based on re-use of elements from schema already represented in Resource Description Framework (RDF), irrespective of the local schema used for the legacy data. An alternative methodology is now being developed. It is based on representing the local schema in RDF and mapping legacy data directly to it, on a one-to-one basis, resulting in lossless publication of linked data. The local RDF elements can then be mapped to elements from other, more general schema using ontological relationships. These mappings can be processed by automatic semantic reasoner software to publish the local linked data to the global infrastructure. The presentation will describe the new methodology using examples taken from experiments with MARC 21 by Metadata Management Associates, and a local relational database schema used by National Library of Scotland for its digital objects collections.

Introduction/background

The final report of the W3C Library Linked Data Incubator Group "examines how Semantic Web standards and Linked Data principles can be used to make the valuable information assets that library create and curate — resources such as bibliographic data, authorities, and concept schemes — more visible and re-usable outside of their original library context on the wider Web."[i] A key recommendation of the report is that "library leaders identify sets of data as possible candidates for early exposure as Linked Data and foster a discussion about Open Data and rights." The report notes that the quantity of bibliographic description data available as Linked Data is relatively small; fewer bibliographic datasets have been published as Linked Data than value vocabularies and element sets. A separate report from the Group defines a "dataset" as a collection of structured metadata describing things of interest, equivalent to a set of library records. A "value vocabulary" is a controlled list of allowed values for a metadata element, and includes authority files, thesauri, and other knowledge organization systems. An "element set" defines the classes and attributes used to describe things of interest. A dataset therefore uses an element set as its structure, and a value vocabulary as some of its content. In the Semantic Web, all three must be formatted as three-part statements, known as triples, using the syntax of Resource Description Framework (RDF) to expose linked data. The three parts represent the Subject-Predicate-Object structure of description logic. To make the statements available for machine processing, the Subject and Predicate must consist of a globally unique identifier, the Universal Resource Identifier (URI). The Object may also be a URI, or a literal string of characters; if it is a URI, it can be matched to the Subject of another triple, to form chains of linked data. Matching the URIs of a Subject produces clusters of linked data with a common subject, similar to a library record. The same triple format can also be viewed as specifying a relationship, the Predicate, between two things, the Subject and Object.

Examples of triple statements in an element set are "This element – is a – Property", "This property – is a sub-Property of – That property", and "This property – has definition – 'The predominate colour of a resource'". Examples from a value vocabulary are "This concept – has notation – 'k'", "This concept – has preferred label – 'blue'", and "This concept – has alternative label – 'turquoise'". And examples from a bibliographic dataset are "This manifestation – has predominate colour – 'k'", "This manifestation – has owner – My Library", and "My Library – has city of location – 'Edinburgh'".

The task set by Incubator Group for library leaders is, in essence, to publish local structured metadata as global linked data in the Semantic Web. The aim is to allow users inside the local environment of the library to benefit from data outside of that environment, and for users outside the local environment to benefit from the library's data. Using the dataset examples given above, a local library user might want to find blue things in art galleries, or a non-local, that is, global, user might find it useful to know if a resource was located in a city they planned to visit.

"Classic" approach

An experiment in publishing library records as linked data was started by the British Library in early 2012. Triples serialized or manifested in RDF/XML format for the British National Bibliography (BNB) are available for download under an open license, Creative Commons Zero (CC0).[ii] The BNB records from which they are derived are encoded in MARC 21 format, so the data needs to be migrated from this format to RDF classes and properties. This is achieved using a map which pairs combinations of MARC 21 tags, indicators, and subfields with single RDF properties. For example, the RDF property "subject" taken from the Dublin Core terms (***dct***) element set[iii] has a mapping from MARC 21 tag 600[iv] for any first indicator, any second indicator excluding "0" and "2", and all subfields. The excluded indicators have mappings to qualified versions of the property which specify Library of Congress Subject Headings or Medical Subject Headings as the associated value vocabulary. The contents of the subfields are concatenated into a single literal string and output as the Object value of a dataset triple using the property as its Predicate. The URI of the Subject of each dataset triple is derived from the record, so a cluster of triples with the same Subject constitutes a linked data version of the original record.

An example dataset triple is:

bnbRes:008739385 dct:subject bnbCon:person/lcsh/HaigDouglasSir1861-1928 .

This triple is serialized or formatted in "turtle", terse triple language (ttl). The three parts of the triple are separated by spaces, and the triple ends with a full-stop. Each URI has two parts: a namespace abbreviation or QName, separated by a colon (:) from a unique local part. Table 1 lists the QNames used in this paper. Each full URI can be reconstituted automatically by substituting the base domain for the QName and removing the separator colon.

This triple states that a resource, in this case a book entitled "Douglas Haig", has as a subject a thing labelled "Haig, Douglas, Sir, 1861-1928", in this case a person. The title of the book and the fact that the resource is a book, and name of the person and that fact that the URI refers to a person, are given by other triples.

Table 1: QNames for namespaces

Abbreviation	Base domain	Namespace
bibo	http://purl.org/ontology/bibo/	Bibliographic Ontology
blt	http://www.bl.uk/schemas/bibliographic/blterms#	British Library terms
bnbCon	http://bnb.data.bl.uk/id/concept/	British National Bibliography concepts
bnbRes	http://bnb.data.bl.uk/id/resource/	British National Bibliography resources
dct	http://purl.org/dc/terms/	Dublin Core terms
dodEl	http://nlsdata.info/dod/elements/	National Library of Scotland Digital Object Database element set
dodRes	http://nlsdata.info/dod/instances/description/	National Library of Scotland Digital Object Database resources
dodWho	http://nlsdata.info/dod/instances/who/	National Library of Scotland Digital Object Database agents
dbp	http://dbpedia.org/resource/	dbPedia: structured information from Wikipedia
geo	http://sws.geonames.org/	Geonames
isbd	http://iflastandards.info/ns/isbd/elements/	International Standard Bibliographic Description
m21	http://marc21rdf.info/	MARC 21 Bibliographic format
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema
skos	http://www.w3.org/2004/02/skos/core#	Simple Knowledge Organization System

The maps used for producing triples for books and serials from the BNB MARC 21 records have been published as RDF templates: British Library Data Model – Book[v] and British Library Data Model – Serial[vi]. Each template shows the properties that are used to generate the dataset triples from a single record. They are taken from several bibliographic namespaces such as ***bibo***, ***dct***, and ***isbd***, as well as a local element set, ***blt***. The local element set contains properties for mapping internal

relationships for aggregations like the publication statement, and for filling gaps not covered by external namespaces, including the provision of inverse properties for swapping the Subject and Object of instance data triples.

The templates also indicate a "range" for each property, expressed by relating the instance URI of the Object of the property to a class. For each instance, these have to be explicitly included in the dataset because they cannot be deduced from the property itself.

Such mappings from a relatively local schema to a schema designed to operate in a global linked data environment can be “lossy”; semantic information may be lost in the transformation of data. This is because the local attribute must have the same or narrower meaning as the global property to maintain semantic coherency. In this case MARC 21, although it is in widespread international use, is intended for more local application than, say, Dublin Core. For example, the mapping of the MARC 21 Uniform Title attribute to the DC Title property removes the information that the content relates to a uniform title, rather than any other kind of title such as parallel or variant title. For this reason the process is often referred to as "dumbing-down" the data. The degree of dumbing is determined by the difference in definition and context between the local and global element; it is minimal when both definition and context are the same, but that is often open to interpretation.

Local element sets

To avoid losing local information in the global Semantic Web, we can represent the local schema as an RDF element set, with comprehensive and unambiguous definitions, and ensure that documentation describes the context for which it is intended.

This is not always easy to achieve. For example, the BNB needs an element set for the MARC 21 schema, but MARC 21 has “messy” semantics that are difficult to define, mixed up with the syntax of tags, indicators, and subfields that constitute the encoding format. The Library of Congress, which maintains the MARC 21 standard, is engaged in the long-term BIBFRAME project for developing a successor. However, Metadata Management Associates (MMA), which maintains the Open Metadata Registry (OMR) has used it to publish a core subset of the MARC 21 Bibliographic format as a series of element sets. They are distributed over several sub-folders of a single base domain (***m21***), each corresponding to a whole or part MARC 21 tag block, as shown in Figure 1. This is necessary to break up the over 14000 distinct properties created for possible combination of tag, indicator, and subfield encoding value. The source of many of these distinctions is the use of "non-filing" indicator values to specify the number of characters to be ignored when using the string value of the first subfield of the tag. There are 9 indicator values for each combination of the indicator with the other indicator and each subfield in the tag, as shown in the partial list of Figure 2.

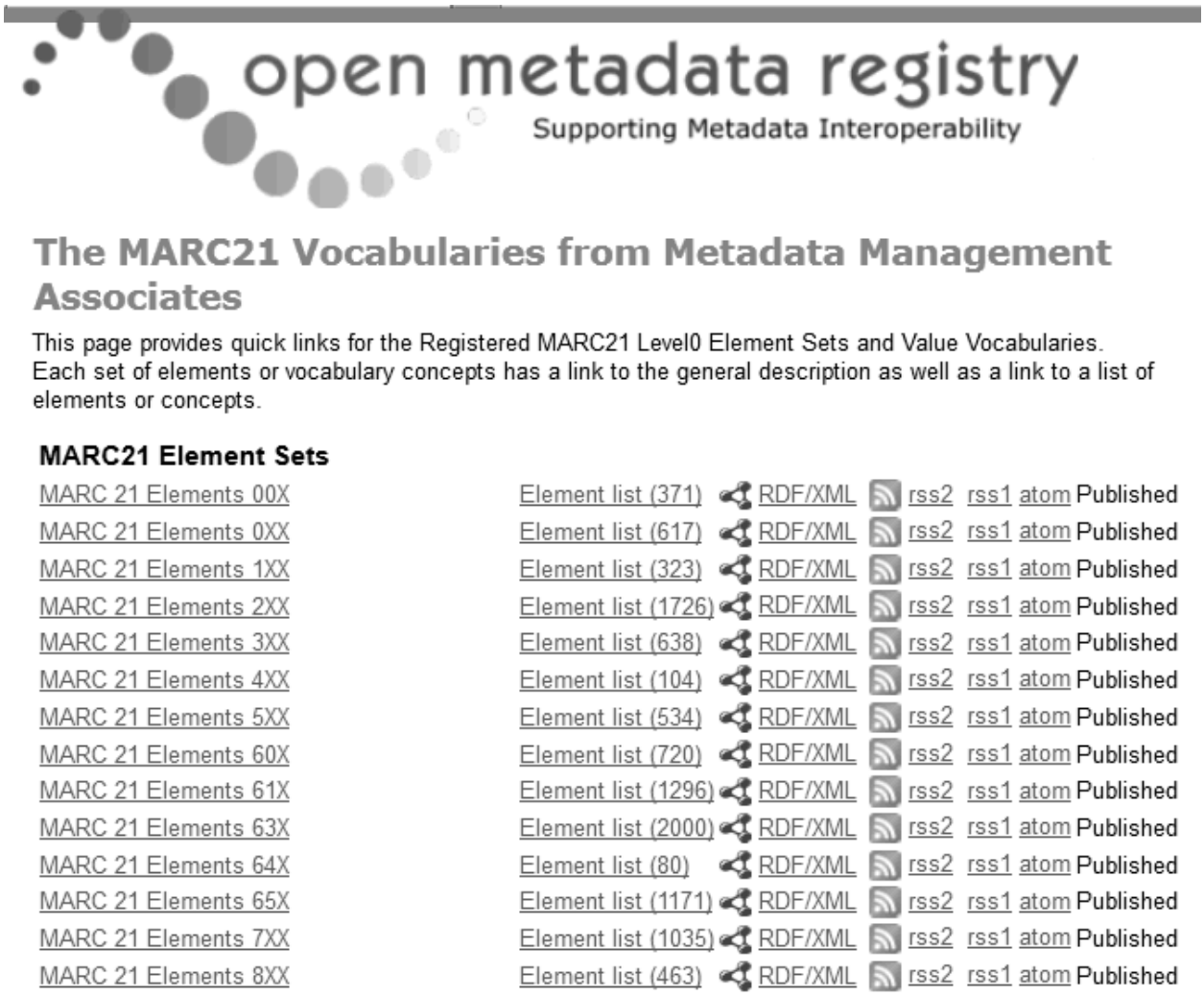

Figure 1: Partial screenshot of the MARC 21 namespace in the Open Metadata Registry.

Figure 2: Partial screenshot of properties based on non-filing characters.

This approach creates an element set at the finest possible level of granularity, and is intended for use with maps for aggregating subfields into statements at the tag level, and for relating the MARC 21 elements to other RDF element sets. Those maps remain in the research phase, so do not offer a short term solution to the problem.

These complexities may not exist, however, in very local bibliographic record schema. These are unlikely to require the breadth or depth of coverage of larger schema like MARC 21, and will often use relational database technologies.

Case: National Library of Scotland

National Library of Scotland (NLS) maintains a Digital Object Database (DOD) of metadata for its local digital library. The database supplies bibliographic metadata for its Digital gallery of digitized materials from its collection, including the First World War official photographs.[vii] An example item is titled "Field Marshall Sir Douglas Haig with H.R.H. Prince Yurihiti of Higashi Fushimi".[viii] The data are stored in normalized form in a relational database management system. Data about a specific entity are stored as values in a row in a table with columns representing the attributes of the entity type or class. Each row has an identifier that is unique within the table. Links between specific entities of different types are stored as pairs of identifiers in a linking table representing the generic relationship between the two entity classes. The descriptive metadata for the example item are stored in a table named "tDescription", with links to entries in other tables containing metadata about related entities, for example "tWho" for people and corporate bodies. This related table is the equivalent of a simple name authority file. Figure 3 shows part of the structure of the database table for descriptive metadata.

The descriptive metadata has the identifier "74548976", and the value for the "description" column starts "Prince Yurihito of Higashi-Fushimi (1867-1922) with Field Marshall Douglas Haig, at a train stop near the Western Front ..." It is linked to an entry in the "tWho" table for Douglas Haig, with identifier "3692" via a linking table which also records the specific sub-relationship, in this instance the subject of the photograph rather than its creator, curator, or owner. Other authority tables cater for subject keywords such as "Trains (vehicle groupings)" and places such as "Europe, Belgium, Flanders (region)".

Figure 3 shows part of the structure of the database table for descriptive metadata.

Name	Type
descriptionTableID	int(11)
objectIdentifierValue	int(11)
title	longtext
shortTitle	varchar(255)
description	longtext
shelfmark	varchar(50)
voyagerBibID	int(11)
isItInTheManuscriptDatabase	tinyint(4)
page	varchar(100)
height	float
width	float
transcription	longtext
notes	longtext
digitisationStatus	varchar(1)
useForCarousel	tinyint(4)
isPageTurnerObject	tinyint(4)

Figure 3: Part of the structure of the database table for descriptive metadata.

It is straightforward to create an RDF element set to represent the semantics of the database structure. A table becomes an RDF class, and a table column becomes an RDF property with the table class as its domain. For example, the DOD's "tDescription" table can be represented as the class "Resource", and the "description" column as the property "has description". The relationships and sub-relationships of linking tables are also represented as RDF properties, with the linked table classes as domain or range respectively, depending on the semantic direction of the predicate. For example, the subject or "depicted" sub-relationship between the "tDescription" and "tWho" tables can be represented as the property "has who depicted" with domain class "Resource" and range class "Who".

Element Sets: Show detail for National Library of Scotland Digital Object Datab

Detail | Elements | History | Maintainers

Label	Type	URI
has place	property	.../dod/elements/tPlace
has place created	property	.../tPlaceType-Placecreated
has place depicted	property	.../tPlaceType-Placedepicted
has place GIS geographic level	property	.../tPlace-gisGeographicLevel
has place GIS place	property	.../dod/elements/tPlace-gisPlace
has place GIS x coordinate	property	.../tPlace.gisXCoordinate
has place GIS y coordinate	property	.../tPlace.gisYCoordinate
has place label	property	.../dod/elements/tPlace-place
has place latitude	property	.../dod/elements/tPlace-latitude
has place longitude	property	.../dod/elements/tPlace-longitude
has shelfmark	property	.../elements/hasResourceShelfmark
has short title	property	.../tDescription-shortTitle
has title	property	.../elements/tDescription-title
has Voyager bib id	property	.../hasResourceVoyagerBibId
has who	property	.../dod/elements/tWho
has who depicted	property	.../elements/tWhoType-Depicted
has who label	property	.../dod/elements/tWho-who
is date or event of	property	.../dod/elements/isDateOrEventOf

Figure 4: Part of the element set of the National Library of Scotland Digital Object Database.

This standard design pattern has been used to develop an element set in the OMR.[ix] Figure 4 shows part of the element set for the NLS DOD. The URIS of the elements are derived directly from the DOD database table and column names. For example, the column name "shortTitle" of the table "tDescription" is used to generate the local part of the URI "tDescription-shortTitle". The element

labels are based on reconverting the CamelCase of the column name and making it a verbal phrase to indicate the direction of the property.

The element set is used to publish parts of the DOD as linked open data. This remains in an experimental stage; the data triples are being used to develop:

- a demonstrator display interface to the linked data for the First World War photographs
- mappings to value vocabularies in external namespaces such as dbPedia (***dbp***), geonames (***geo***), and the BNB
- mappings to element sets in external namespaces such as ***dct***
- links to external data triples such as the Canadian "Out of the trenches" project[x]

Examples of mappings from a DOD property to a ***dct*** property are:

1. ***dodel:tDescriptive-title rdfs:subPropertyOf dct:title.***
2. ***dodEl:tWhoType-Depicted rdfs:subPropertyOf dct:subject .***

The mapping or ontological property used in both examples is taken from the RDF Schema (RDFS) namespace. The *subPropertyOf* property is associated with a semantic inference rule which can be used to generate or entail new statements:

> *If property P1 is a sub-property of property P2 (P1 rdfs:subproperty P2);*
> *And P1 is the predicate of a data triple (Subject P1 Object);*
> *Then a data triple with the same subject and object and P2 as the predicate is entailed (Subject P2 Object).*

For example, using the first mapping, the DOD data triple:

dodRes:D74548976 dodEl:t tDescriptive-title "Field Marshall Sir Douglas Haig with H.R.H. Prince Yurihiti of Higashi Fushimi" .

entails a new data triple:

dodRes:D74548976 dct:title "Field Marshall Sir Douglas Haig with H.R.H. Prince Yurihiti of Higashi Fushimi" .

Similarly, using the second mapping, the DOD data triple:

dodRes:D74548976 dodEl:tWhoType-Depicted dodWho:D3692 .

entails the triple:

dodRes:D74548976 dct:subject dodWho:D3692 .

The original triple says that the example photograph depicts Douglas Haig, and the entailed triple says that the photograph has the subject Douglas Haig.

An example of a mapping from a DOD value to a BNB value is

dodWho:D3692 skos:exactMatch bnbCon:person/lcsh/HaigDouglasSir1861-1928 .

This triple states the DOD URI and the BNB URI refer to the same entity, the Douglas Haig depicted in the photograph and the subject of the book titled "Douglas Haig". The triple can be included in an application which finds books published in Britain about the same person, organization, event, etc. depicted in the NLS photograph collection. Information from Wikipedia can be linked using the mapping triple:

dodWho:D3692 skos:exactMatch dbp:Douglas_Haig,_1st_Earl_Haig .

All such links from the local element set and its data to external element sets and linked data can all be expressed in machine-readable triples.

Advantages of local RDF element set

There are several advantages in creating and using a local element set:

- Published linked data loses no information, because it uses properties which exactly match the local schema.
- Other communities can see the semantics and structure of the local data schema expressed in RDF. They can determine where the linked data comes from, its context and how it was formed.
- Other communities can re-use the schema for their own local applications, and to map from their own local schema, although this can, of course, be lossy.
- Anyone can map the local element set to other element sets such as ***bibo***, ***dct***, ***isbd***, etc.
- You can have your cake, and eat it! The local published linked data can be consumed by semantic processes for use in other applications, including dumbing down, but it always remains in its original state for further consumption at the local and global level.

A recipe for how to have your cake and eat it!

1. The local organization publishes its local schema as an RDF element set and value vocabularies.
2. The local organization publishes its local data triples using the local RDF vocabularies.
3. Anyone publishes mappings from the local element set to other, more global elements.
4. Anyone entails or publishes mapped global data triples using “reasoner” software.

Conclusion

Publishing linked data from local legacy records without losing information requires the development and publication in RDF of the metadata schemas used to store and maintain the records, unless those schemas completely conform to a more global schema that is already published in RDF.

The development and maintenance of RDF element sets and value vocabularies representing local schemas may require significant resources, depending on the complexity and coherency of the schemas. Smaller schemas, and in particular those based on normalised relational database structures, require fewer resources, and their data is often describe unique local resources. An open infrastructure which includes software, documentation, and services for to support vocabulary maintenance, mapping, and entailment, is evolving that can support the lossless publication of linked data for these resources.

References

[i] Library Linked Data Incubator Group Final Report: W3C Incubator Group Report 25 October 2011. Available at: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

[ii] British National Bibliography (BNB) - Basic Representation. Available at: http://datahub.io/dataset/bluk-bnb-basic

[iii] DCMI Metadata Terms. Available at: http://dublincore.org/documents/dcmi-terms/

[iv] MARC 21 Bibliographic – Full: 600 - Subject Added Entry-Personal Name (R). October 2007. Available at: http://www.loc.gov/marc/bibliographic/bd600.html

[v] British Library Data Model – Book. V.1.4 August 2012. Available at: http://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf

[vi] British Library Data Model – Serial. V.1.0 August 2012. Available at: http://www.bl.uk/bibliographic/pdfs/bldatamodelserial.pdf

[vii] National Library of Scotland. First World War 'Official Photographs'. Available at: http://digital.nls.uk/first-world-war-official-photographs/

[viii] Field Marshall Sir Douglas Haig with H.R.H. Prince Yurihiti of Higashi Fushimi. Available at: http://digital.nls.uk/first-world-war-official-photographs/pageturner.cfm?id=74548976

[ix] Element Sets: Show detail for National Library of Scotland Digital Object Database elements. Available at: http://metadataregistry.org/schema/show/id/64.html

[x] Out of the trenches: a linked open data project. Available at: http://www.canadiana.ca/en/pcdhn-lod