Linked data for manuscripts in the Semantic Web

Gordon Dunsire

Summer School in the Study of Historical Manuscripts, Zadar, Croatia, 26–30 September 2011; Topic II: New Conceptual Models for Information Organization

Pre-print

Published by Gordon Dunsire

Edinburgh 2012

LINKED DATA FOR MANUSCRIPTS IN THE SEMANTIC WEB

Gordon Dunsire

Independent consultant, Edinburgh, UK

Abstract

This paper briefly describes the basic concepts of Resource Description Framework (RDF), the basis of linked data in the Semantic Web. It then discusses how the elements of bibliographic metadata are identified and represented for machine-processing within the RDF environment, and the current status of elements specific to the cataloguing of manuscripts. The paper describes in detail a simple methodology for creating metadata in the form of RDF statements or triples from existing bibliographic records, using examples from manuscript descriptions. The paper concludes with a number of questions about the interaction of the manuscripts community with the Semantic Web.

Keywords

Resource Description Framework (RDF), linked data, Semantic Web, manuscript cataloguing, bibliographic namespaces, legacy metadata

Basic concepts of the Semantic Web and Resource Description Framework

The Semantic Web is the "web of linked data" envisioned by the W3C to "enable people to create data stores on the Web, build vocabularies, and write rules for handling data" and to "enable computers to do more useful work ".[1] Computers are networked on a global scale, operate 24 hours a day and seven days a week, and can process data in suitable formats much faster than a human being. The Semantic Web therefore needs a standard machine-processable format for stored data, vocabularies, and data-handling rules. The format recommended by the W3C is Resource Description Framework (RDF),[2] which stores data as simple, single statements known as triples, and supports rule-based processing. The ability to process rules allows a form of logical reasoning or inferencing, the "semantic" in the Semantic Web. While computers cannot determine the veracity of a single triple, they can detect incoherency in a set of triples using such rules.

A triple is so-called because it requires each statement to be in three parts: the subject of the statement, the nature or aspect of the subject, and a value for that aspect. The three parts are usually referred to, in order, as subject, predicate, and object. To all intents and purposes, this represents metadata or data about data; every triple has a subject which the statement is "about". For example, the simple statement "The title of this manuscript is 'Ode to himself'" can be represented by a triple with the subject "This manuscript", the predicate "has title", and the object value "Ode to himself": "This manuscript has title 'Ode to himself'", which has the same meaning as the original statement.

Other examples of subject-predicate-object triples are ”This letter” + “has author” + “Jane Doe”, and “This codex” + “has material” + “papyrus”.

There is a need for an unambiguous way of identifying the parts of a triple to allow efficient machine-processing. The labels used by humans, such as “This codex” and “has title” are not suitable because they are often ambiguous. Humans frequently refer to the same thing with different labels, and to different things with the same label. For example, does “title” refer to a designation of nobility, or the title of a manuscript, and does the latter include assigned titles and translated titles?

RDF requires the use of a Uniform Resource Identifier (URI) to identify components of a triple. A URI can be any combination of numbers and letters, provided it is unique. There is no intrinsic meaning to a URI; it is just a machine-readable identifying label. Human-readable labels such as “title” or “has title” can be associated with the URI for a particular predicate by using it as the subject of other triples such as “PredicateURIX” + “has label” + “title” or “PredicateURIX” + “has label” + “titre” or “PredicateURIY” + “has label” + “title of nobility”.

The existing utility of the Web’s Uniform Resource Locator (URL) can be exploited to create URIs. A URL is already machine-readable with a regular syntax, and is unambiguous even at global scale. The combination of numbers and letters in a URI can look like a URL, for example “http://iflastandards.info/ns/isbd/elements/P1004”, but it does not in principle lead to a Web page or other online document. There are advantages, however, in using a so-called http URI or “cool URI”.[3] RDF requires the subject and predicate of a triple to be URIs, while the object can be a URI or a literal string such as “Ode to himself”. When the object of a triple is a URI, it is possible to match it to the subject of another triple, allowing the triples to be chained into “linked data”.

The W3C Library Linked Data Incubator Group identified three categories of RDF representations.[4] A “dataset” is a collection of structured metadata describing things, such as manuscripts in an archive. A “value vocabulary” is a set of defined values such as a controlled terminology that can be used as the object of a triple in a dataset, for example a name authority file or list of subject headings. An “element set” is a set of defined properties and classes that can be used in a dataset, value vocabulary, or another element set to describe entities, attributes, and relationships of interest.

Identifying bibliographic metadata

It is necessary to assign URIs to the specific resources described by records in catalogues and finding aids, such as manuscripts, collections, digital surrogates, etc. These will be the subjects and objects of triples in a dataset, and require URIs. Vocabularies, authority files, subject headings, classifications,

and other knowledge organization systems (KOS) need to be represented as value vocabularies so that their URIs can be the objects of triples. Triples with object URIs can form linked data, but triples with literal URIs cannot and effectively terminate a chain of one or more triples. Many controlled vocabularies used in bibliographic metadata have been published as value vocabularies, including Library of Congress Subject Headings (LCSH), French subject headings in Rameau, German subject headings in Schlagwortnormdatei (SWD), the Dewey Decimal Classification, and the many vocabularies in RDA: resource description and access. Mappings and other relationships between terms in different vocabularies have also been represented as linked data, for example the mappings between English, French, and German subject headings developed by the Multilingual Access to Subjects (MACS) project.[5]

It is also possible and appropriate to represent the attributes and relationships found in bibliographic schemas as RDF properties, each of which has its own URI, in an element set. These properties can then be used as the predicates in triples. Schemas published or being developed in RDF by library standards bodies include RDA, International Standard Bibliographic Description (ISBD), and Functional Requirements for Bibliographic Records (FRBR) and related models.[6] More information about the processes involved is available for RDA,[7] the Functional Requirements models,[8] ISBD,[9] and UNIMARC.[10]

Entities in schemas are usually represented in element sets as RDF classes, which define sets of things with common characteristics. In many cases it is possible to represent sub-types of entity either as a sub-class, a more specific property, or a term in a vocabulary; for example, “author” as a sub-type of “creator” can be represented as a sub-class of Creator, as the property “has author”, or as the term “author” in a vocabulary of creative agents. Each approach has different advantages and disadvantages which need to be considered when using an element set to create a dataset of bibliographic linked data.

Schemas have differing purposes. The Functional Requirements family of models analyse the attributes of, and relationships between, entities which support sets of user-centred tasks. ISBD provides a record structure and content standard for the exchange of national bibliographic metadata. UNIMARC is an encoding standard for ISBD records and authority records based on Functional Requirements for Authority Data (FRAD). There is significant overlap in the entities, attributes, and relationships identified by each standard, but there has been little or no opportunity to avoid duplication in the RDF representations because of the differing stages of development and lack of a coordinating infrastructure. This is unlikely to be a significant issue in the future because alignments between standards can also be represented in RDF for machine-processing.

A namespace is a set of URIs with the same common root or "base domain" managed with a single infrastructure. Each element set of RDF classes and properties and each value vocabulary usually has its own namespace; for example "http://iflastandards.info/ns/isbd/terms/contentform/" is the base domain for the namespace for the ISBD content form vocabulary. URIs are created by adding a "local part" to the base domain; the local part must be locally unique, and the base domain must be globally unique. For example "http://iflastandards.info/ns/isbd/terms/contentform/" plus the local part "T1009" gives the URI "http://iflastandards.info/ns/isbd/terms/contentform/T1009" for the content form labelled "text".

The overlap between bibliographic standards results in the same, or very similar, properties in different namespaces; for example, both ISBD and RDA have properties for "title proper". This overlap extends to other namespaces relevant to bibliographic metadata, including Dublin Core and the Bibliographic Ontology. This results in a choice of which URI to use in a triple, which will be influenced by the context of the namespace and the relationship between the class, property, or value and other elements in the namespace and, indeed, other namespaces.

Identifying manuscripts

There are, as yet, no published namespaces exclusively intended to for application to manuscripts and their collections.

The general bibliographic namespaces contain elements and values which pertain to manuscripts as well as other forms of information resource, such as the property "title" and term "paper". In addition, some namespaces include elements that are specific to manuscripts. For example, there are over 40 properties[11] and 8 value vocabulary concepts[12] with "manuscript" in their labels in the Open Metadata Registry.

The importance of representing metadata for manuscripts in RDF is exemplified by the development of Collex, "an open-source collections- and exhibits-builder designed to aid humanities scholars working in digital collections or within federated research environments like NINES [Nineteenth century Scholarship Online]".[13] Collex uses RDF "[t]o make the relationships between objects more explicit" and allow users to add their own descriptive tags to objects.[14]

Methodology for creating linked data from bibliographic records for manuscripts

Very large numbers of bibliographic records exist, but not as linked data. It is very difficult to estimate just how many. The largest union catalogue, OCLC's WorldCat, has several hundred million

records in its database, but these have been contributed by a small minority of the world's libraries and archives. It is very likely the total number of records is in the billions. The number of distinct information resources described by those records is also unknown, as a single resource may be the focus of multiple duplicate records.

Most of the metadata contained in such records is of high quality relative to other sources of metadata, such as social networking websites ranging from Wikipedia to local reader circles. The metadata has usually been created by highly trained and experienced cataloguers and indexers using standards developed over decades. Each record may generate many triples, perhaps 30 from a single MARC record. There are therefore potentially very, very large numbers of triples currently locked inside catalogues and finding-aids. Releasing this data to the Semantic Web will provide raw material for the development of applications and services taking advantage of the utility of linked data, as well as promote professional standards and practices.

From record to triples in 9 stages

The following methodology produces linked data triples from existing metadata records. It can be adapted to the creation of triples from scratch, and can be applied to any type of information resource, including manuscripts. No special software tools are required, and it is possible to carry out the process manually using a text editor. However, large-scale processing would require the development of specific programs and workflows to extract and map the data from multiple records, as for example in the projects involving the British National Bibliography.[15]

Step 1: Take a record

Linked data requires a description of a resource to be presented as a set of statements, with each statement giving a value which describes a specific aspect of the resource. This allows a single triple to represent each statement by storing the value as its object and the aspect as its predicate or property. The subject of every triple is the same: the resource itself.

For example, a description of a scientific manuscript might include information about its title, who wrote it and when, what subjects are covered, what material is used, and so on. This information needs to be laid out as a set of pairs of attributes and their values, as shown in Table 1.

Field/attribute	**Value**
Record ID	54321
Title	Notes on an electrical experiment
Author	Michael Faraday
Date	1845

LCSH	Impedance (electricity)
Material	Paper
Content form	Text

Table 1. Attributes and their values for a simple description of a manuscript

This is usually the way that information is stored as fields in, or available for export from, a machine-readable database. A unique value is required to act as an identifier for the set of statements. It will also be used as the identifier for the resource itself. The record identifier used in a machine-readable database is a useful source of the value.

Step 2: Disaggregate to single statements

The attribute/value pairs are reformatted into subject-predicate-object statements by using the record identifier as the subject of each statement, as shown in Table 2.

Record ID	**Attribute**	**Value**
54321	(has) title	Notes on an electrical experiment
54321	(has) author	Michael Faraday
54321	(has) date	1845
54321	(has) LCSH	Impedance (electricity)
54321	(has) material	Paper
54321	(has) content form	Text

Table 2. Description formatted as a set of statements

For each statement, the attribute will form the predicate and the value will be the object. It helps to clarify the statement by making the predicate a verbal phrase. This can often be done by adding a simple possessive verb to the attribute label. For example, the last row of Table 2 can be stated as "The resource described by record 54321 has [the] content form Text".

Each statement can now be represented as an RDF data triple.

Step 3: Create URI for record

The subject of a triple must be a URI. Therefore it must be globally unique, and not used to identify any other resource. In this example "54321" may be unique to the database providing the record, but it cannot be guaranteed to be unique outside of that local environment.

Fortunately, the infrastructure which supports URLs for web pages and other documents can be used to generate a URI from a record identifier. The domain names used in URLs are unique – only one person or organization can control a domain at any one time. For example, the domain "www.nsk.hr"

is licensed to the National and University Library in Zagreb, and is the basis of the URL for the Library's homepage: http://www.nsk.hr/.

A cultural heritage organization can create a "cool URI" for a resource in its collections by appending the corresponding record ID from the catalogue to a special domain, or a unique sub-folder of a general domain; for example, the special domain "MyCollectionX.com" plus "54321". Adding the "http://" prefix allows additional functionality, to give "http:// MyCollectionX.com /54321" as the URI. This is not a URL, and using it in a standard web browser will display a "page not found" error unless the organization has arranged for additional processing of the URI as if it were a URL.

Step 4: Replace record ID with URI

The record ID can now be replaced with the resource URI in the set of statements, as shown in Table 3.

Resource URI	**Attribute**	**Value**
mlx:54321	(has) title	Notes on an electrical experiment
mlx:54321	(has) author	Michael Faraday
mlx:54321	(has) date	1845
mlx:54321	(has) LCSH	Impedance (electricity)
mlx:54321	(has) material	Paper
mlx:54321	(has) content form	Text

Table 3. Statements with the subject replaced by a URI

The resource URI is given as a compact URI (CURIE).[16] The first part, "mlx:", is an abbreviation for "http:// MyCollectionX.com/". This makes it easier for human application developers to refer to the URI, and the abbreviation is automatically expanded to obtain the full URI when processed by a computer program. The resource URI in the example can therefore be considered to be from the "My collection X" namespace, or just the "mlx" namespace. The namespace or base domain abbreviation is not fixed and each developer can choose their own, but ad hoc preferences are evolving as usage increases, especially with the schema namespaces discussed below.

Step 5: Find URIs for attributes

The predicate in each statement about the resource is represented by an RDF property. RDF requires that the predicate in a triple is a URI, so the next step is to find a URI for a property matching the attribute in each statement.

Some bibliographic metadata communities have represented their schemas in RDF, with properties based on schema attributes and relationships. The URIs assigned in these representations usually have

a base domain common to the schema; that is, they are managed in a namespace in the same way as the URIs for individual resources in the example collection. For example, such element set namespaces have been published for Dublin Core terms (dct), ISBD (isbd), FRBR (frbrer), RDA (rda…), Bibliographic Ontology (bibo), and others.

Attributes can be directly matched to RDF properties if they are taken from a schema that has published properties in its own namespace. For example, if a manuscript record was created using the consolidated edition of ISBD, then the attribute “Dimensions” has a corresponding property “has dimensions” in the ISBD element set namespace, with the URI “http://iflastandards.info/ns/isbd/elements/P1024”.

If the source of an attribute is not known, or the schema does not have an RDF namespace, then the attribute should be matched to an equivalent property in a published element set namespace. In order to minimise loss of information, a property with the same, or nearly similar, definition as the attribute should be chosen. In order to avoid semantic incoherency, the property needs to have the same or broader definition as the attribute. If the property has a narrower definition, there is a possibility that values assigned to the attribute in any specific record will lie outside of the definition, or meaning or semantic, of the property.

There is no requirement that attributes used in a record must be matched to properties from the same namespace. Instead, it is possible to match attributes to properties from different namespaces. This can be very useful when seeking the closest possible equivalent property for each attribute. The aim is to obtain the URI of a property that has the same or broader definition as the attribute.

The first attribute in the example record is “title” or “has title”. There are similar properties in most of the bibliographic element set namespaces because it is an important attribute. In some schemas, it is designated as mandatory, for example ISBD, or “core”, as with RDA.

In fact, some namespaces offer a bewildering number of properties that may match a title attribute. For example, ISBD has nearly 40 properties in its namespace with “title” in the label, ranging from “has title” to “has note on parallel titles and parallel other title information”. Choosing the nearest match involves comparing the definition of the attribute with the definition of the property, and taking into account additional information such as scope, cataloguing rules for the attribute, and context. For example local manuscript cataloguing guidelines may specify a method for assigning a title when the item being described lacks one; this may be very different from the approach taken by international book cataloguing rules. Workflows for choosing an appropriate property are beginning to appear, such as the LODE-BD recommendations for the AGRIS network.[17]

Other namespaces may only have one property for title, for example Dublin Core terms. Such properties will usually have a very broad definition, and are not always suitable if information loss between the statement and the triple is to be minimised. Examples of possible matches to the example record's title attribute are given in Table 4. These are taken from the Dublin Core terms, ISBD, and RDA element set namespaces.

Namespace	Property label	Property definition
dct	title	A name given to the resource.
isbd	has title proper	Relates a resource to the title proper (the chief name of a resource, i.e. the title of a resource in the form in which it appears on the preferred source of information for the resource).
rda	Title proper	The chief name of a resource (i.e., the title normally used when citing the resource).

Table 4. Examples of "title" properties from different namespaces

If we assume that the record is based on ISBD, we can choose the "has title proper" property, with the URI "http://iflastandards.info/ns/isbd/elements/P1014", or compact URI "isbd:P1014".
Note that it is not possible to make such an obvious decision for some of the other attributes from the record because ISBD does not cover attributes such as "author" or "LCSH" or other types of heading or entry point. That is, the record may be based on ISBD, but it is not a purely ISBD record. This is why the ability to choose properties from different namespaces is important.
So there is no ISBD property for the next attribute, "author". The nearest equivalent properties from the Dublin Core terms and RDA namespaces are given in Table 5.

Namespace	Property label	Property definition
dct	creator	An entity primarily responsible for making the resource.
rda	author	A person, family, or corporate body responsible for creating a work that is primarily textual in content, regardless of media type (e.g., printed text, spoken word, electronic text, tactile text) or genre (e.g., poems, novels, screenplays, blogs). Use also for persons, etc., creating a new work by paraphrasing, rewriting, or adapting works by another creator such that the modification has substantially changed the nature and content of the original or changed the medium of expression.

Table 5. Properties corresponding to the "author" attribute

In this instance, the RDA property is likely to be closer in meaning to the "author" attribute than the Dublin Core terms property, because the RDA property has a similar label, and "primarily textual in content" in the definition narrows the focus. The URI of the RDA property is

“http://rdvocab.info/roles/author” or “rdaroles:author” in compact form. The remaining attributes, with the exception of “LCSH”, can be matched to RDF properties in similar ways, as shown in Table 6.

Attribute	Property definition	Property compact URI
(has) date	Relates a resource to the date on which it is officially offered for sale or distribution to the public, usually given in the form of a year.	isbd:P1018
(has) base material	The underlying physical material of a resource.	rdagr1:baseMaterial
(has) content form	Relates a resource to a category that reflects the fundamental form or forms in which the content is expressed.	isbd:P1001

Table 6. Attributes matched to property URIs

Note that match of the “date” attribute requires more information than available in the property definition. This information is found in the text of the consolidated edition of ISBD.

The attribute “LCSH” requires some additional processing. Library of Congress Subject Headings is a controlled vocabulary of subject topics, so the attribute can be considered to be a sub-attribute of “(has) subject”. Other sub-attributes used in other records might include “Dewey Decimal Classification” or “Art and Architecture Thesaurus”. In RDF, the vocabulary used for the values of a general attribute is implicit in the URI for a specific value, if a URI has been published. LCSH is one of several such “value vocabularies” that have been represented in RDF, so a URI can be found for all specific basic topics, as discussed below.

The “LCSH” attribute can therefore be represented by a more general “subject” attribute and matching property. A suitable property is available in the Dublin Core terms namespace, with the definition “The topic of the resource” and compact URI “dct:subject”.

Step 6: Replace attributes with URIs

The attributes in the statements derived from the example record can now be replaced by the matching property URIs, as shown in Table 7.

Subject URI	Attribute property URI	Value
mlx:54321	isbd:P1014	Notes on an electrical experiment
mlx:54321	rdarole:author	Michael Faraday
mlx:54321	isbd:P1018	1845
mlx:54321	dct:subject	Impedance (electricity)

mlx:54321	rdaGr1:baseMaterial	Paper
mlx:54321	isbd:P1001	Text

Table 7. Statements with the attribute replaced by a property URI

Step 7: Find URIs for values

If object of a triple is a URI rather than a literal value, then a machine can link it to the subject URI of another triple. This is the basic idea of linked data in the Semantic Web.

Values in the record which are taken from controlled vocabularies may have URIs. If the vocabulary has a representation in RDF, each term will have a URI, required for the subject of triples storing data about each term, such as its definition, and preferred and alternate labels or names, etc.

Bibliographic authority files traditionally control the labels used as headings for "authors", including artists, performers, directors, and other roles, and subject topics and classification notations. More recently, RDA has introduced controlled vocabularies for many other bibliographic attributes, and ISBD uses controlled terms in its area 0 for content form and media type. URIs for the values in the example record may therefore exist for the author, subject, material, and content form statements. They are unlikely to exist for the title and date statements, as these are not usually controlled. There is an RDF representation of the Virtual International Authority File (VIAF).[18] Searching for "Michael Faraday" leads to the heading "Faraday, Michael, 1791-1867". This has the URI, of "http://viaf.org/viaf/38158158", displayed as a "Permalink" on the VIAF website. This can be given the compact URI "viaf:38158158".

Library of Congress Subject Headings has an RDF representation available from the Library of Congress Authorities & Vocabularies service.[19] The URI for the subject heading "Impedance (electricity)" is given as "http://id.loc.gov/authorities/subjects/sh85064610". A compact form is "lcsh:sh85064610".

RDA has a controlled vocabulary for base materials, with an RDF representation available from the Open Metadata Registry.[20] The concept "paper" has the URI "http://rdvocab.info/termList/RDAbaseMaterial/1011", which can be shortened to "rdabm:1011". The ISBD vocabulary for content form also has an RDF representation in the Open Metadata Registry. The concept "text" has the URI "http://iflastandards.info/ns/isbd/terms/contentform/T1009", which can be given as the compact URI "isbdcf:T1009". It is worth noting that the same URI has preferred labels for "text" in Spanish, as "texto", and in Croatian, as "tekst".

Step 8: Replace values with URIs

The values of the statements corresponding to the example record can now be replaced with URIs where available, as shown in Table 8.

subject	**predicate**	**object**
mlx:54321	isbd:P1014	“Notes on an electrical experiment”
mlx:54321	rdarole:author	viaf:38158158
mlx:54321	isbd:P1018	“1845”
mlx:54321	dct:subject	lcsh:sh85064610
mlx:54321	rdaGr1:baseMaterial	rdabm:1011
mlx:54321	isbd:P1001	isbdcf:T1009

Table 8. Set of RDF triples derived from statements

Each statement has been reformulated as an RDF triple, with a subject and predicate URI, and an object URI or literal. Literals are enclosed in double quotes.

Step 9: Publish triples (linked data)

The final step is to publish the set of RDF triples derived from the example record. Triples can be stored and displayed in a number of formats, called serializations. One of these uses terse triple language (ttl), which makes it easier to see the three-part structure of each triple. If compact URIs are used, the set of triples must be preceded by a declaration of the abbreviations used, so that the proper URI can be reconstituted automatically. The ttl serialization of the example triples is shown in Figure 1.

```
@prefix dct: <http://purl.org/dc/terms/> .
@prefix isbd: <http://iflastandards.info/ns/isbd/elements/> .
@prefix isbdcf: <http://iflastandards.info/ns/isbd/terms/contentform/> .
@prefix lcsh: <http://id.loc.gov/authorities/subjects/sh85041643> .
@prefix mlx: <http://.../> .
@prefix rdabm: <http://rdvocab.info/termList/RDAbaseMaterial/> .
@prefix rdaGr1: < http://rdvocab.info/elements> .
@prefix rdarole: <http://rdvocab.info/roles/> .
@prefix viaf: <http://viaf.org/viaf/> .
mlx:54321 isbd:P1014 “Notes on an electrical experiment” .
mlx:54321 rdarole:author viaf:38158158 .
mlx:54321 isbd:P1018 “1845” .
mlx:54321 dct:subject lcsh:sh85064610 .
mlx:54321 rdaGr1:baseMaterial rdabm:1011 .
mlx:54321 isbd:P1001 isbdcf:T1009 .
```

Figure 1. TTL serialization of a set of triples

The triples can now be used to link to other triples. A human-readable version of the original example record can be reconstituted using triples from the value vocabularies which link to the preferred labels for the concepts. For example, a triple in the ISBD content form vocabulary has the URI for “text” concept as its subject, with a URI for the property “preferred label” as its predicate, and the literal “text” as its object. The property is itself taken from the Simple Knowledge Organization System (SKOS) element set. Similarly, VIAF uses a property from the Friend of a Friend (FOAF) element set to link the URI for a person to the authorized form of their name, while LCSH uses a property from the Metadata Authority Description Schema (MADS) to link the URI for a subject to the authorized form of heading.

Another way of displaying triples is as a graph, in which nodes represent the subject and object of a triple, with a connecting line representing the predicate. A round node is used for a subject or object URI, while a rectangular node displays an object literal. The connecting line uses an arrow to indicate the direction from subject to object. The RDF graph for the reconstituted record is shown in Figure 2.

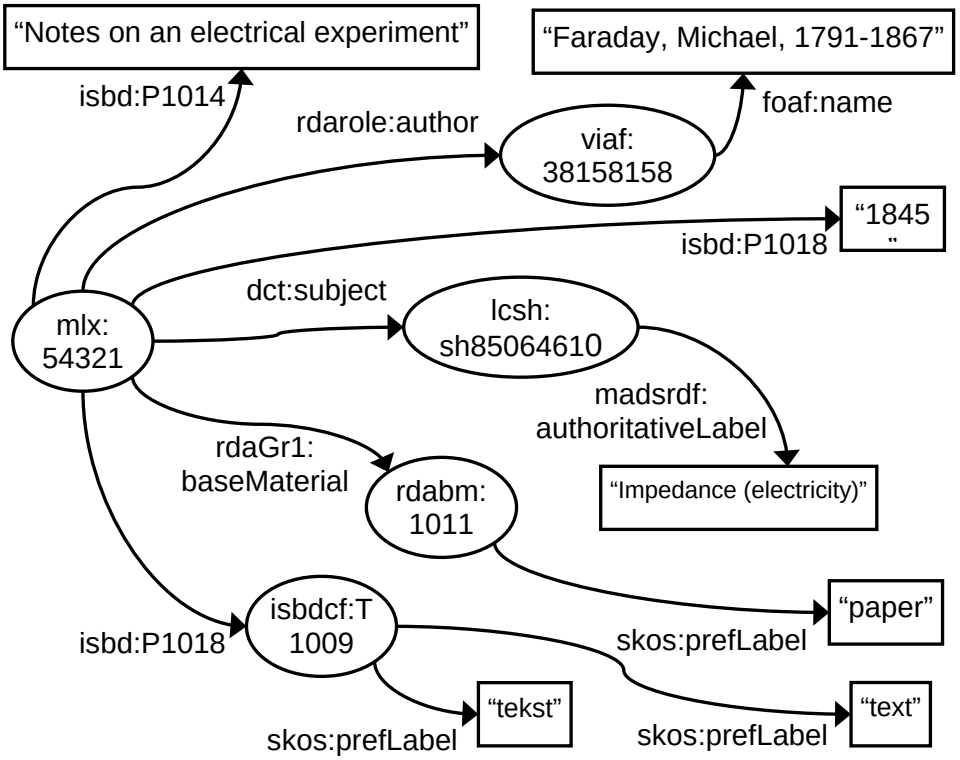

Figure 2. RDF graph of linked data corresponding to an example record for a manuscript

The graph includes the Spanish and Croatian translations of the ISBD content form, showing how useful the linked data approach can be: these terms can be used by Spanish and Croatian speakers to find the manuscript described.

Any of the URI (round) nodes in the graph can be linked to a node in another graph or set of triples by matching the URI. For example the manuscript graph could be linked to the graph of a digitized version or a commentary by using properties based on entity relationships. Ultimately, the Semantic Web may consist of a single, very large graph connecting all information resources.

Conclusion

The utility of metadata for manuscript resources can be improved by representation in RDF as linked data for the Semantic Web. There is currently no widely-available infrastructure focussing specifically on manuscripts, but components of element sets and value vocabularies for general bibliographic resources are likely to cover many of the specific needs of manuscript scholarship and curation. It is possible to create linked data from scratch, or convert it from existing metadata records, using small-scale manual methods.

This paper leaves a number of questions for further study, research, and application:

- Is there a need for new RDF classes and properties to describe aspects of manuscripts that are not covered by general library and archive element sets?
- Is it more efficient and effective to create RDF triples describing manuscripts from scratch, or by creating the metadata in existing record structures and then converting it?
- What tools are required to create, store, and publish triples for manuscripts more effectively?
- What kinds of systems are needed to integrate manuscript metadata into semantic resource discovery services?
- Will a semantic view of metadata for manuscripts lead to changes in cataloguing and indexing practices, and what training will be required for manuscript cataloguers and indexers?

References

[1] W3C Semantic Web. Main page. Last modified: 2012 [cited: 2012-04-09]. Available at: http://www.w3.org/2001/sw/wiki/Main_Page
[2] W3C. RDF Working Group. Resource description framework (RDF). 2004 [cited: 2012-04-09]. Available at: http://www.w3.org/RDF/
[3] W3C. Cool URIs for the Semantic Web. W3C Interest Group note 03 December 2008 [cited: 2012-04-09]. Available at: http://www.w3.org/TR/cooluris/
[4] W3C. Library Linked Data Incubator Group. Datasets, value vocabularies, and metadata element sets.

W3C Incubator Group Report 25 October 2011 [cited: 2012-04-09]. Available at: http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

[5] MACS: Multilingual Access to Subjects. Last updated: 2011 [cited: 2012-04-09]. Available at: http://www.nb.admin.ch/nb_professionnel/projektarbeit/00729/00733/index.html?lang=en

[6] Dunsire, Gordon; Mirna Willer. Standard library metadata models and structures for the Semantic Web. // Library Hi Tech News 28, 3(2011), 1-12. Available at: http://www.emeraldinsight.com/journals.htm?articleid=1926531&show=abstract [cited: 2012-04-09].

[7] Hillmann, Diane; Karen Coyle; Jon Phipps; Gordon Dunsire. RDA vocabularies : process, outcome, use. // D-Lib Magazine 16, 1/2(2020). Available at: http://www.dlib.org/dlib/january10/hillmann/01hillmann.html [cited: 2012-04-09].

[8] Dunsire, Gordon. Interoperability and semantics in RDF representations of FRBR, FRAD and FRSAD. // Concepts in context : proceedings of the Cologne Conference on Interoperability and Semantics in Knowledge Organization, July 19th-20th, 2010 / edited by Felix Boteram, Winfried Gödert and Jessica Hubrich. Würzburg : ERGON, 2011. Pp. 133-147.

[9] Willer, Mirna; Gordon Dunsire; Boris Bosancic. ISBD and the Semantic Web. // JLIS.it: Italian Journal of Library and Information Science 1, 2(2010). Available at: http://leo.cilea.it/index.php/jlis/article/view/4536 [cited: 2012-04-09].

[10] Dunsire, Gordon; Mirna Willer. UNIMARC and linked data. // IFLA Journal 37, 4(December 2011), 314-326. Earlier version available at: http://conference.ifla.org/sites/default/files/files/papers/ifla77/187-dunsire-en.pdf [cited: 2012-04-09].

[11] See: http://metadataregistry.org/schemaprop/search?sq=manuscript&commit=Search+Element+Sets [cited: 2012-04-09].

[12] See: http://metadataregistry.org/conceptprop/search?concept_term=manuscript&commit=Search+Vocabularies [cited: 2012-04-09].

[13] Collex [cited: 2012-04-09]. Available at: http://www.nines.org/about/software/collex/

[14] Knight, Kim. Collex : research report. 2006 [cited: 2012-04-09]. Available at: http://transliteracies.english.ucsb.edu/post/research-project/research-clearinghouse-individual/research-reports/collex

[15] Wilson, Neil. Establishing the connection : creating a linked data version of the BNB. 2011 [cited: 2012-04-09]. Available at: http://www.slideshare.net/nw13/establishing-the-connection-creating-a-linked-data-version-of-the-bnb

[16] W3C. CURIE Syntax 1.0: a syntax for expressing compact URIs. W3C Working Group note 16 December 2010 [cited: 2012-04-09]. Available at: http://www.w3.org/TR/curie/

[17] Subirats Imma; Marcia Zeng. Report on how to select appropriate encoding strategies for producing Linked Open Data (LOD)-enabled bibliographic data. [Version 1.1] [cited: 2012-04-09]. Available at: http://aims.fao.org/lode/bd

[18] VIAF: Virtual international authority file. © 2010-2011 [cited: 2012-04-09]. Available at: http://viaf.org/

[19] Library of Congress. Authorities and vocabularies [cited: 2012-04-09]. Available at: http://id.loc.gov/

[20] Metadata Management Associates RDA base material. Last updated 2009 [cited: 2012-04-09]. Available at: http://metadataregistry.org/vocabulary/show/id/35.html