Reconsidering Universal Bibliographic Control in Light of the Semantic Web

By Gordon Dunsire, Diane Hillmann, and Jon Phipps

Introduction

The goal of universal bibliographic control (UBC) as a world-wide system for the control and exchange of bibliographic information acknowledges the resource discovery metadata requirements of modern, global scale users of information. The first decade of this millennium has seen a significant change in thinking about the functions of UBC and how they can best be realized.

This is illustrated by the case of VIAF (Virtual International Authority File). VIAF links together authority records from multiple national and regional agencies to form clusters of data labeling the same person, corporate body, or other entity. The service developed after the abandonment of attempts by the International Federation of Library Associations and Institutions (IFLA) to establish a framework that ensured that each entity of interest was referenced by a single, unambiguous name, "authorized heading", or similar label, to be used by all agencies creating bibliographic metadata. Decisions on what form the label should take proved difficult to make when local users require displays conforming to their cultural norms for language, script, and form. Initiatives by IFLA and other organizations to replace the single human-readable label with a language- and script-independent numerical identifier have also failed in the context of top-down UBC [Tillett, 2008].

Although new and on-going projects such as the International Standard Name Identifier (ISNI) continue such endeavors, there is still no general agreement amongst the world's cataloguing communities to prefer any single system or type of identifier. Yet VIAF "expands the concept of [UBC] by … allowing national and regional variations in authorized form to coexist; and … supporting needs for variations in preferred language, script and spelling" [VIAF]. VIAF achieves this by retaining the authority data supplied by each agency and treating it equally; it is allowed to speak for itself. As well as offering locally preferred forms and local identifiers, the service also offers an identifier for the cluster of matched headings. This can act as the identifier for an aggregation node in a UBC system of linked identifiers; there is no need to provide a human-readable UBC label for a cluster because local forms are available.

VIAF shows that UBC for bibliographic authority data can be achieved from the bottom up, from local to global, by linking identifiers for local data to an aggregator identifier without transforming or discarding the local data. We believe that this approach can be extended to other types of bibliographic metadata by using a wider range of linkages to relate and map the data. In particular, we have been looking at the potential of the distributed architecture and "reasoning" capabilities of the Semantic Web to provide a future for UBC.

The Resource Description Framework and the Semantic Web

During the last few years, interest and understanding of the role of the Resource Description Framework (RDF) and the Semantic Web environment based on it have begun to percolate through the library community. Among those who speak and write about these technologies and changes in thinking they represent, Karen Coyle has been among the most visible and prominent. Her recent publications for Library TechSource are particularly helpful for those looking for solid background reading on these issues from a library point of view [Coyle, 2010]. The RDF primer provides a good

technical introduction [RDF, 2004]. In essence, RDF is a language read not by people but by applications and it is read in such a way as to not lose information.

In contrast to the record-based approach of traditional library metadata, in the RDF data model the focus is on individual metadata statements represented by three-part data triples in the form subject-predicate-object; for example "This website (subject) has a creator (predicate) with the name 'Smith' (object)". To allow unambiguous manipulation by machine, RDF requires each subject and predicate to be a Uniform Resource Identifier (URI); the object can also be a URI, or it can be a string of characters known as a "literal". Linking the object of one statement to the subject of another, via URIs, results in a chain of linked statements, or linked data. This avoids the ambiguity of using natural language strings as headings to match statements. As a result, a literal object terminates a linked data chain, and literals are generally used for human-readable display data such as labels, notes, names, etc. A set of URIs assigned to specific RDF properties and classes, using a single management infrastructure is called a "namespace". The RDF approach is very different from the traditional library catalogue record exemplified by MARC21, where descriptions of multiple aspects of a resource are bound together by a specific syntax of tags, indicators, and subfields as a single identifiable stream of data that is manipulated as a whole. In RDF, the data must be separated out into single statements which can then be processed independently from one another; processing includes the aggregation of statements in to a record-based view, but is not confined to any specific record schema or source for the data. Statements or triples can be mixed and matched from many different sources to form many different kinds of user-friendly displays.

The data content of a statement is kept separate from its semantic content, which is expressed through one or more statements using ontological predicates, or RDF properties. The bottom-up view of a bibliographic record becomes a set of data triples with the same subject URI using classes and properties from bibliographic element namespaces. The bottom-up view of a metadata schema is a set of ontological triples, or RDF ontology, using the same classes and properties as their subject URIs. Thus the ontology triples are about the classes and properties used in data triples about bibliographic entities.

RDF is immediately relevant to UBC because an RDF triple actually stores a linkage or relationship between two identifiers or an identifier and a label, and the identifiers are designed for global scale. Because the property used for the relationship has its own URI, it can be linked or related to another property as well as an RDF class or literal. This "property of properties" or ontological property mechanism is very useful for relating properties from namespaces for different metadata schemas.

Among the ideas Coyle and others have articulated is that the ability to associate different data triples from various sources and using different metadata schema is one of the most important benefits of this new environment.

> "It's an unfortunate fact that many systems combine data from different sources using only the 'dumb down' method, reducing the metadata to the few matching elements and resulting in the least rich metadata record possible. This results in a tremendous loss of data and an inferior user experience. The 'smart up' method uses all or most of the data from the different sources, resulting in enhanced information. For example, the Open Library record is able to link to any number of information sources both from its pages for books and its pages for authors, in part because it can store linkable data from any source without having to be concerned about fitting that data into a particular record format. It also means that it can create a display that is richer than any one data source." [Coyle, 2009 1)

The notions of 'dumbing down' and 'smartening up' are critical in understanding the differences between top-down and bottom-up views of data relationships and mapping. Rethinking the notion

of UBC in this environment requires that both viewpoints be kept in mind when defining relationships as well as maps.

> "Despite the fact that some metadata developers do not think of the relationships as a form of mapping, the DCMI 'dumb-down' principle is an example of a pre-defined semantic mapping that retains great value in a mapping environment, particularly when the level of granularity between schemas is very different. One of the core principles behind the notion of dumb-down is a requirement that the more refined of two related properties must be considered to be a subset (subproperty) of the more broadly defined property. Even though the loss of specificity may mean less clarity of understanding, the data does not lose meaning when 'dumbed down' to its broader relative. Of equal and perhaps even greater value, broader properties defined by Simple DC may be 'smartened up' to provide increased specificity when possible." [Dunsire, 2011 1]

The RDF environment provides basic ontological properties that are used by software "reasoners" to make logical inferences. This effectively results in the automatic generation of new data triples, a mechanism that provides interoperability of data when an RDF ontology is based on a map between properties and classes from two or more metadata schemas.

Linked data: interoperability from the bottom up

Most of the major bibliographic metadata schemas used by libraries are at some stage in the process of representation in RDF as properties and classes [Dunsire, 2011 2]. Real examples are used in the following discussion.

The diagrams use RDF graphs to represent triples. An arc represents the predicate or RDF property of a triple, connecting two elliptical nodes representing the URIS of the subject and object. The arrow of an arc points to the object. An object which is a literal is shown as a rectangle pointed to by only one arc; values without identifiers cannot form merged nodes in an RDF graph and therefore terminate the linked data chain. Arcs and nodes have labels based on their URIs. For the sake of readability, URIs are given in modified Compact URI (CURIE) format, with the namespace in an abbreviated form called a qname separated by a colon from the associated English label in quotes. For example, the URI http://iflastandards.info/ns/isbd/elements/P1004 has the CURIE isbd:P1004, modified for display on a graph as the label isbd:"has title proper".

Some frequently-used properties are abbreviated further:

d (domain)	= rdfs:domain
eP (equivalent property)	= owl:equivalentPropertyOf
r (range)	= rdfs:range
sC (subclass)	= rdfs:subClassOf
sP (subproperty)	= rdfs:subPropertyOf

The qnames used are:

dc	= http://purl.org/dc/elements/1.1/
dct	= http://purl.org/dc/terms/
ex	(Example instance namespace)
frbrer	= http://iflastandards.info/ns/fr/frbr/frbrer/
isbd	= http://iflastandards.info/ns/isbd/elements/
owl	= http://www.w3.org/2002/07/owl#

rdact = http://rdvocab.info/termList/RDACarrierType/
rdafrbr = http://rdvocab.info/uri/schema/FRBRentitiesRDA/
rdagrp1 = http://rdvocab.info/Elements/
rdaopen (RDA unconstrained by FRBR; under construction)
rdfs = http://www.w3.org/2000/01/rdf-schema#
skos = http://www.w3.org/2004/02/skos/core#

Diagram 1 shows the RDF graph of a set of triples using properties representing the RDA core attributes for a manifestation. Most triples have the same subject URI, ex:1, and a literal object. The exception is the object of the carrier type which is linked to an RDA controlled vocabulary represented in RDF/SKOS. It is worth noting that the publication statement is an aggregated statement [Hillmann, 2010]. It can be replaced with or broken down into a more refined set of triples with attribute properties such as place of publication, publisher's name, and date of publication, some of which can be linked to further triples.

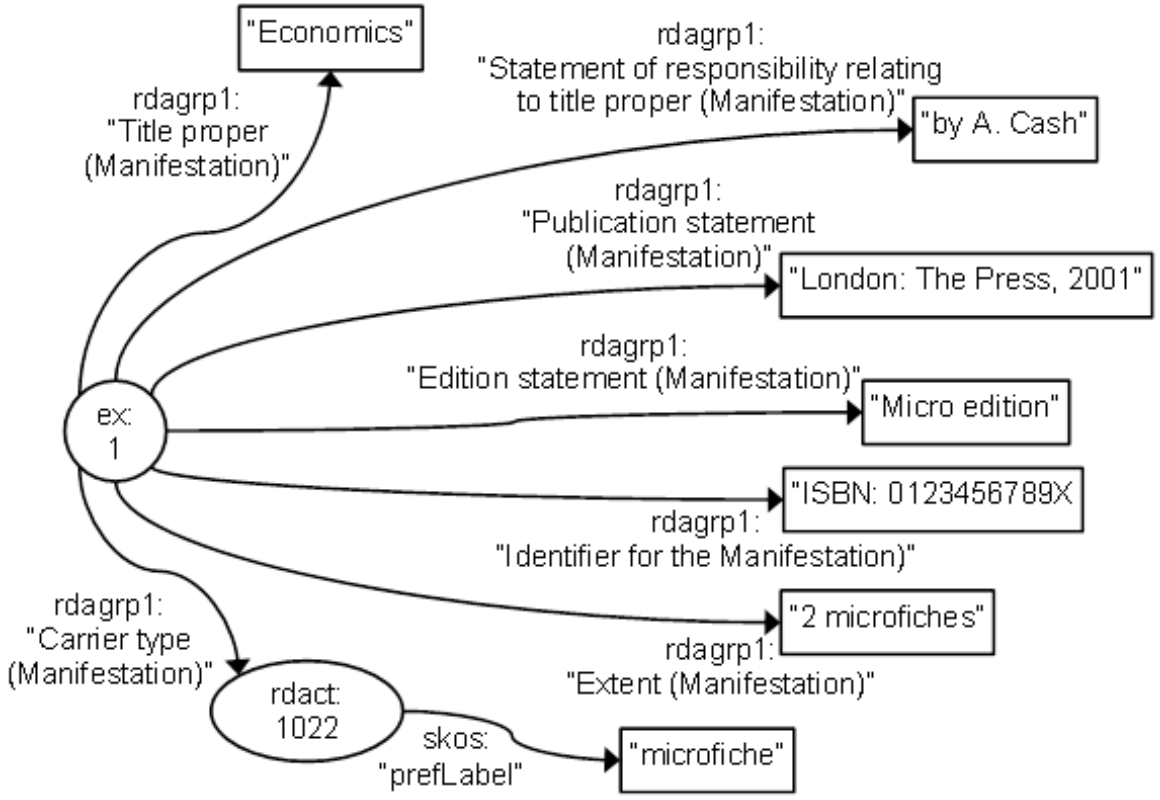

Diagram 1: RDF graph of core attributes of a specific RDA manifestation

The graph in Diagram 1 contains more information than the specific instance data represented in its triples. Each property has a URI which is itself the subject of a set of triples specifying properties of that property, including a human-readable label. The label "Title proper (Manifestation)" is such a triple; another literal with the same subject is the definition "The chief name of a resource (i.e., the title normally used when citing the resource)." A property can also be related to another property or

class via a triple using various special ontological properties. Of particular interest to mappings are properties from the RDFS namespace with labels “domain”, “range”, “subclass of”, and “subproperty of”. Diagram 2 is the RDF graph of the set of ontology triples of the RDA property labeled “Title proper (Manifestation)”. It contains triples which say that this property is a subproperty of two other RDA properties, each of which is a subproperty of yet a fourth RDA property. It also says that the original property and one of the related properties have a domain of the RDA class labeled “Manifestation”. This ontology is derived from an RDA entity-relationship diagram and the addition of general versions of the RDA properties [Hillmann, 2010].

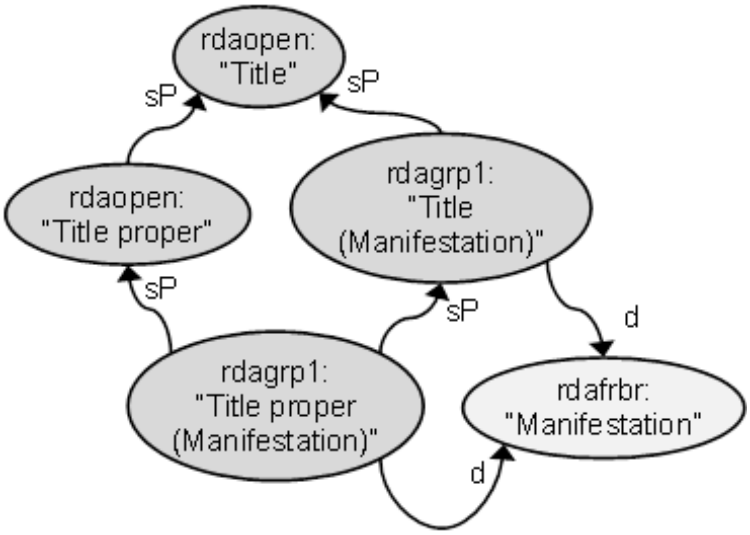

Diagram 2: RDF graph of the ontology of rdagrp1: “Title proper (Manifestation)

The important feature for mapping is that ontological triples can be used to infer, or entail, new data triples. For example, a data triple using a property which has a domain entails another data triple saying that the subject is a type or member of the class specified as the domain.

For example, the resource identified by the URI of the subject of a triple using the “Title proper (Manifestation)” property can be inferred to be an instance of the class “Manifestation” because the domain of the property is specified as that class. In other words, the statements: “<This resource> <has title proper (manifestation)> ‘A Title’ ”, and the property "<has title proper (manifestation)> <has domain> ‘manifestation’ ” imply the statement “This resource is a manifestation”. This is important because, from an RDF point of view, inferring that a resource is a type of thing, and directly stating that it is a type of thing, are fundamentally equivalent. This differs from XML where, unless you explicitly declare a resource to be a type of thing, you can't know what type of thing it is. This has great value in metadata aggregation because it allows a resource to be described from the perspective of many different communities as many, usually similar, types of thing.

The “range” property similarly entails that the object of a data triple using the property is an instance of the class specified as the range. The “subclass” property entails that a URI which is an instance of the subclass is also an instance of the super-class. If “This is a Border collie” and “’Border collie’ is a subclass of ‘Dog’”, then “This is a dog”.

The “subproperty” property is very useful for the interoperability of data from different metadata schemas. Using a property which has a “subproperty” property entails a triple using the specified super-property with the same subject and object. We can state that a property is a subproperty of a second property if it is determined that the definition of the first property is entirely subsumed by the definition and meaning of the second property. In Diagram 2, “Title proper (Manifestation)” is a subproperty of “Title (Manifestation)” because a title proper is a type of title. So a data statement

using the “Title proper (Manifestation)” property, say “This resource has title proper (manifestation) ‘That title’” entails the new statement “This resource has title (manifestation) ‘That title’”.

Diagram 3 is an RDF graph of the triples that can be entailed from the “Title proper (Manifestation)” triple in Diagram 1 using the ontology in Diagram 2. The new triples are shown with dashed line arcs. Any of the triples can be used in any other linked data graph. Entailed triples can be generated once and stored as instance data triples, or generated on the fly as an application requires them.

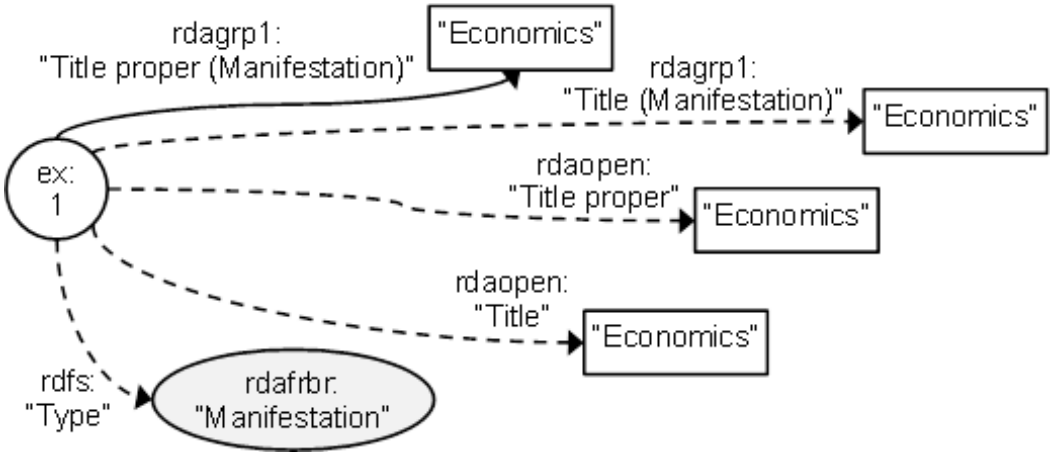

Diagram 3: RDF graph of entailed triples

The same RDF ontological properties can be used to link properties and classes from different namespaces, allowing two ontologies to be mapped and combined into a super-ontology.

Diagram 4 shows the RDF graph of the ontology of the ISBD property labeled ”has title proper”. It is a subproperty of another ISBD property, and both have the ISBD class “Resource” as a domain.

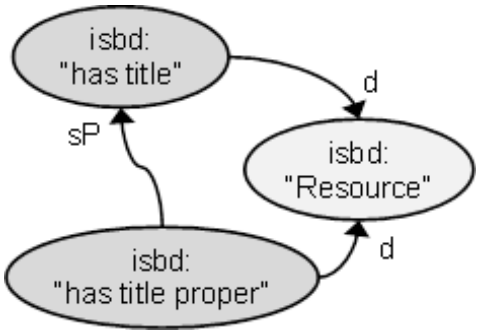

Diagram 4: RDGF graph of the ontology of isbd “has title proper”

A combined ontology for the RDA and ISBD “title proper” properties can be developed if ontological relationships can be determined between any of the RDA properties in Diagram 2 and any of the ISBD properties in Diagram 4. This requires careful analysis of the property definitions to determine if they are equivalent or if one subsumes the other. Semantic incoherence must be avoided in any entailments that result from linking the ontologies; for example, the RDA class “Manifestation” is not a subclass of the ISBD class “Resource”, and this should not be entailed inadvertently from the combined ontology.

Diagram 5 is the combined ontology of Diagram 2 and Diagram 4. It identifies two pairs of related RDA and ISBD properties. ISBD’s ”has title proper” is a subproperty of the general RDA property

"Title proper" to avoid linking the classes "Resource" and "Manifestation" and generating a false entailment. ISBD's "has title" is linked to the general RDA property "Title" using the OWL (web ontology language) property "equivalent property", indicating that the definitions, while different, have the same semantic meaning. Thus a data triple using one of the properties entails a triple using the other property, with the same subject and object.

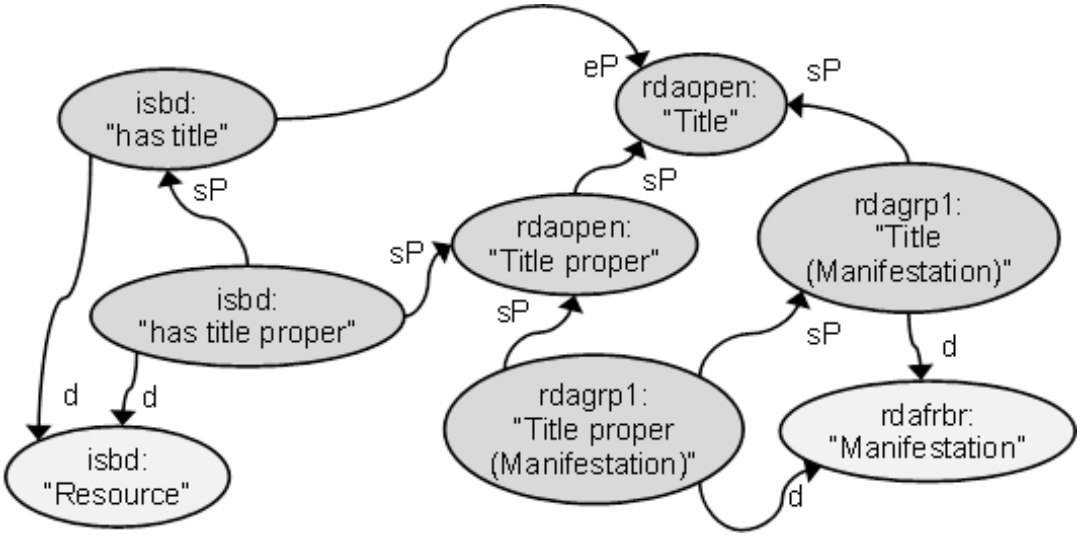

Diagram 5: RDF graph of combined ontology of the ISBD and RDA "title proper" property

This process can be repeated to link properties and classes from any namespace, ontology, or general RDF graph. Diagram 6 is the RDF graph of Diagram 5 combined with properties from the Dublin Core namespaces.

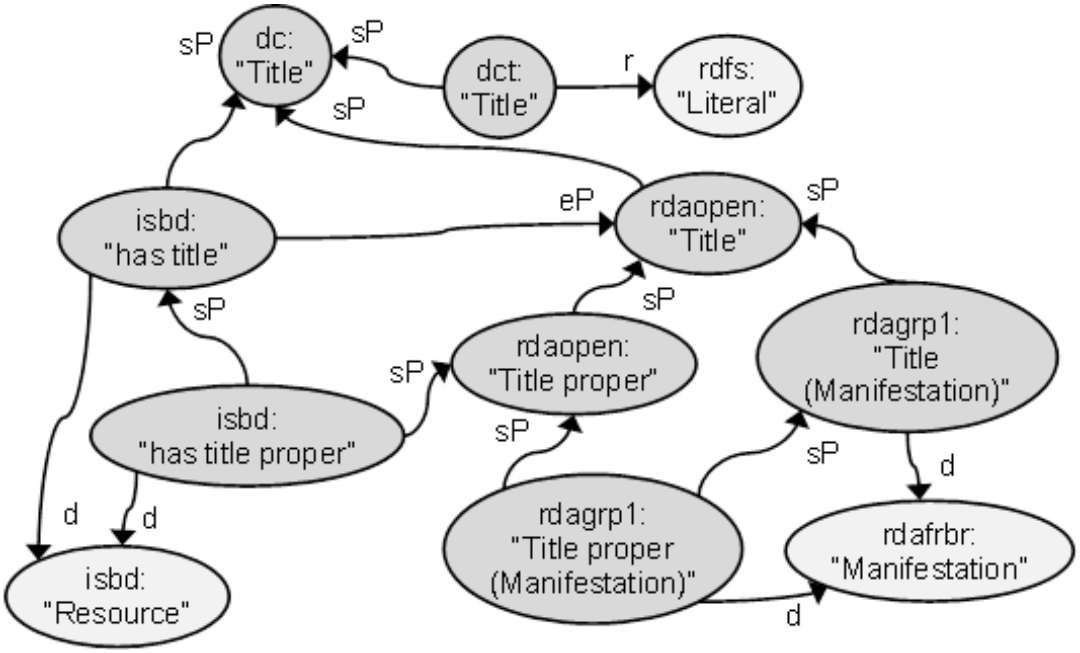

Diagram 6: RDF graph of ontology of ISBD, RDA, and DC "title" properties

These combined ontologies effectively constitute mappings between the attributes, represented as RDF properties, of one metadata schema and another. The mappings are instantiated through entailed instance data triples. Diagram 7 shows RDF graphs of three instance data triples based on RDA, ISBD, and DC Terms properties respectively, with triples for RDA general and DC properties entailed from Diagram 6.

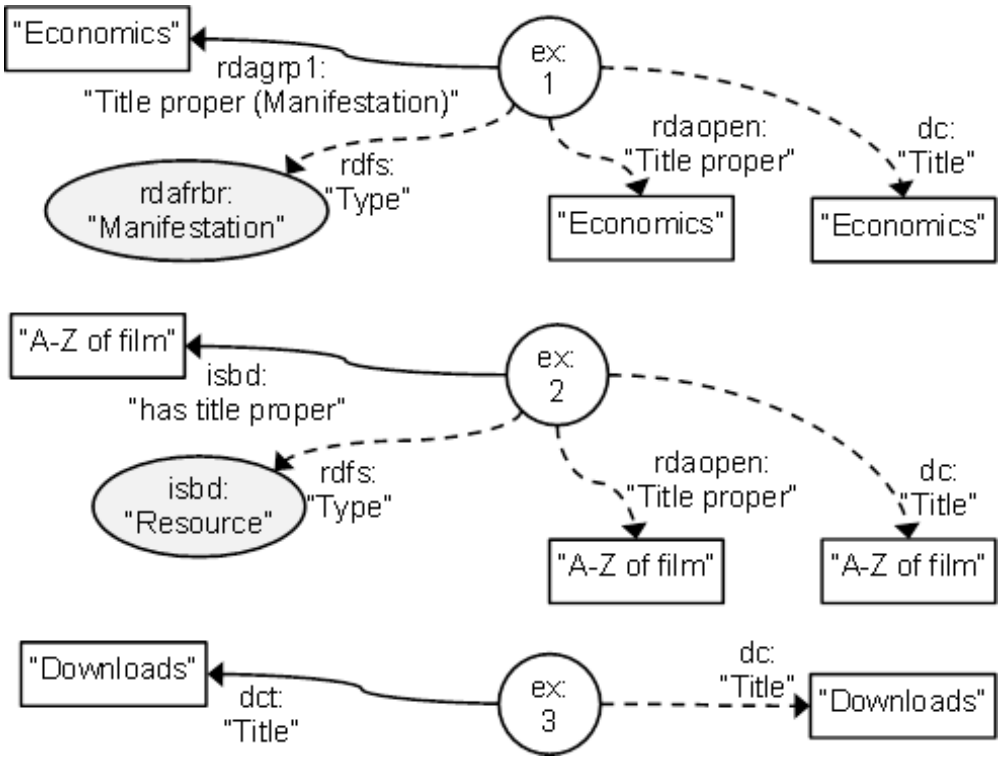

Diagram 7: RDF graphs of RDA, ISBD, and DCT instance triples with selected entailments

These entailed triples can be used by applications generating “title proper” and general title indexes from instance data derived from different metadata schema for a range of types of information resource.

This approach has several important features:

- The original data triples remain intact and available to any application.
- The entailed data triples are available to any application.
- The values of the objects in the original triples remain unchanged in the derived triples.
- Multiple levels of semantic granularity are supported with entailments derived from sub-properties available for all less specific, “dumber” properties.

Functional futures

In the library world, “mapping” has been synonymous with “crosswalking”. However, much of the crosswalking research of the last decade or so focuses less on the creation of the maps and more on the resulting transformational process. In this traditional world, maps are developed, ingested and maintained as documents (usually spreadsheets) that are not necessarily actionable outside of applications. Thus a further, but separate, step beyond the intellectual process of creating a map is the creation of programs that implement the mapping and transform data based on the decisions

made during the creation of the maps. Because of the centralization of the traditional library metadata world, this focus has led to a situation where only a few large organizations with significant development resources have thought much about--or built--crosswalks or programs to use them.

But the meaning and implications of “mapping” changes radically when moving from a database and record based approach to an open, multi-domain, global, shared environment based on linked data technologies -- where anybody can say anything about any topic, validity constraints are not acknowledged, a nearly infinite number of properties can be defined to describe an infinite number of entities, and authority is multi-dimensional and often ephemeral. The classic approach to such apparent chaos is to attempt tighter control over the creation process, more filtering, additional restrictions, and less access. This approach hinders appreciation and use of the broad diversity of perspective that comes with a world of open data.

Mapping through linked data semantics does not have the same result as mapping through crosswalks. Crosswalks are generally managed through applications that not only translate values stored in one metadata schema to another, but also transform the values themselves. In the new environment, without the necessity of considering a transformative end process, the idea of one ‘best’ map (or collection of maps) no longer seems relevant. Limiting mappings to close-match equivalences also seems outdated. A more useful approach would define a mapping strategy that is open, extensible, and built using technology that goes beyond the spreadsheet.

In an RDF-based data environment, transformation can still take place within an application, but as a separate operation from mapping. Transformation performed in a semantically mapped environment removes some of the anxiety surrounding potential degradation when round-tripping the contents of data elements because it is done without loss of information about the original semantics. Without the necessity of defining an “authoritative” or “best” mapping, a metadata element can have more than one set of semantics at the same time; this means it should be a simple matter to move from different but compatible definitions as needed within an application.

RDF permits and encourages the publishing and aggregation of metadata statements rather than records that describe a bibliographic resource. Thus, RDF “records” represent arbitrary collections of statements. These statements may be defined and validated by any number of metadata “formats” and a collection of such statements can be composed of properties selected by the publisher of the metadata. Systems aggregating this published data are free to choose the properties from this record that the system “understands”. Properties defined by MARC21 can be freely mixed with properties defined by RDA, or any other vocabularies. RDF “records” can vary widely in terms of overall content and there are no constraints on what a system may publish. For example, the DC Application Profile framework [Coyle, 2009 2] provides a means of specifying the components of a “description set” to be used to aggregate statements for an application. This is distinct from the creation, validation and storage of resource descriptions. The process of metadata creation benefits greatly from the current notion of bounded and constrained records in order to enforce rational limits on how data is validated and managed.

This bottom-up approach to UBC allows the functions of bibliographic metadata curated by the library community to remain focused on the requirements of that community. Expressing the schemas and data which support those functions in the open framework of RDF linked data allows other communities to re-use the data to meet similar requirements in their own environment. But it also allows the data to function for any purpose imaginable. What those new functions might be, and how well library metadata serves them, should trouble libraries no more than how a patron uses the information from their collections. The true worth of library linked data lies in its consistency and completeness, of high value when it is mixed into the uncontrolled environment of the Semantic Web.

Conclusion

Looking forward and backward as the authors have above is only part of the picture. Libraries are at a period of great change, driven to a large extent by technology. The transition we face as we look toward the future requires that we understand the environment before us, and those librarians with one foot in the world of cataloging and another in the emerging metadata space have particular issues to address. Before we begin making decisions about how our future will look, we need to understand the burdens of our prior modes of thinking, and where they must change. What the term 'bibliographic control' means is a critical place to start.

The lessons of the past and the opportunities facing us combine to suggest strongly that continuing to interpret "bibliographic control" as a monolithic, top-down effort designed to achieve universality--as the library world has traditionally done--is not going to allow us to take advantage of new technologies or new ways of thinking about and building metadata. A new paradigm of bottom-up allows "control as co-ordination" through semantic mappings--making the essential shift from controlling the data to controlling the semantics that will allow us to move forward, taking our legacy data with us.

References

[Coyle, 2009 1] Coyle, K. Metadata mix and match. (2009) *Information Standards Quarterly*, 21(1). Available at: http://kcoyle.net/isqv21no1.pdf

[Coyle, 2009 2] Coyle, K., & Baker, T. (2009) Guidelines for Dublin Core Application Profiles. Available at: http://dublincore.org/documents/profile-guidelines/

[Coyle, 2010] Coyle, K. Understanding the Semantic Web: bibliographic data and metadata. (2010) *ALA Techsource*, 46(1). (an excerpt is available for free at: http://alatechsource.metapress.com/content/p3022442071g7655/fulltext.pdf)

[Dunsire, 2011 1] Dunsire, G., Hillmann, D., Phipps, J. & Coyle, K. (2011). A reconsideration of mapping in a semantic world. *Proc. Int'l Conf. on Dublin Core and Metadata Applications 2011*. Available at: http://dcevents.dublincore.org/index.php/IntConf/dc-2011/paper/view/52/6

[Dunsire, 2011 2] Dunsire, G. & Willer, M. (2011) Standard library metadata models and structures for the Semantic Web. *Library Hi Tech News*, 28(3), 1 – 12. Originally presented at the IFLA Conference in Sweden, August 2010 and available at: http://www.ifla.org/files/hq/papers/ifla76/151-dunsire-en.pdf

[Hillmann, 2010] Hillmann, D., Coyle, K., Phipps, J. & Dunsire, G. RDA vocabularies: process, outcome, use. *D-Lib magazine*, 16(1/2). Available at: http://www.dlib.org/dlib/january10/hillmann/01hillmann.html

[RDF 2004] RDF primer, W3C Recommendation 10 February 2004. Available at: http://www.w3.org/TR/rdf-primer/

[Tillett, 2008] Tillett, B. A review of the feasibility of an International Standard Authority Data Number (ISADN). Prepared for the IFLA Working Group on Functional Requirements and Numbering of Authority Records. 1 July 2008. Available at: http://archive.ifla.org/VII/d4/franar-numbering-paper.pdf

[VIAF] VIAF (the Virtual International Authority File). Available at: http://www.oclc.org/research/activities/viaf/