Caliber 2011

Theme of the Conference: Towards Building a Knowledge Society: Library as Catalyst for Knowledge Discovery and Management

Theme paper "**Web Resource Management and Semantic Web**

Gordon Dunsire
Consultant
11 Cobden Terrace, Edinburgh EH11 2BJ, Scotland

Abstract

The paper discusses the role of libraries in the Semantic Web. It provides a brief introduction to the Semantic Web and its technologies, including Resource Description Framework, and their relevance to library metadata. It describes the evolution of library expertise in developing metadata schemas and creating structured metadata, and the resulting quantities of rich, high-quality metadata records that might provide critical mass to the development and utility of the Semantic Web. The paper outlines modern analyses of user tasks and requirements for knowledge discovery, including International Standard Bibliographic Description, Functional Requirements for Bibliographic Records, and RDA: resource description and access, and how the associated models and rules are being made available to the Semantic Web. Examples of current initiatives to make legacy metadata available are given, and the work of the W3 Library Linked Data Incubator group described. The paper then discusses the potential impact of the Semantic Web on library metadata management, as the focus changes from record to individual statement, and the beneficial impact on users of Web and library services.

Keywords

Libraries, Semantic Web, knowledge organization systems, metadata, information retrieval systems

1. Introduction to the Semantic Web

This paper discusses the potential role of libraries in the Semantic Web. This section introduces relevant concepts and terminology from the Semantic Web.

"The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

"For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning."

"Human language thrives when using the same term to mean somewhat different things, but automation does not."

"... an ontology is a document or file that formally defines the relations among terms."

"More advanced applications will use ontologies to relate the information on a page to the associated knowledge structures and inference rules." [1]

The Semantic Web is based on two fundamental building blocks: Resource Description Framework (RDF) [2], and the Uniform Resource Identifier (URI) [3].

RDF is used to make statements about Web resources in the form of subject-predicate-object expressions, called triples. A triple corresponds to a single statement, for example "The author of this paper is Gordon Dunsire". The subject of a triple is what the statement is about, in this case "this paper". The predicate is the specific property being described: the attribute "author" or role of authorship. The object of a triple is the value of the property in the context of the subject, in this instance "Gordon Dunsire". The triple for our statement is thus "this paper"-"author"-"Gordon Dunsire", and in this form is readily understandable by humans. It cannot effectively be processed by a computer, however. The machine will erroneously treat "this paper" as the same thing even if it refers to another paper, say at the same conference. The machine will treat "author" as a different property to "writer" used in another triple. It cannot distinguish things which have the same name or label but different meanings, or things which have the same meaning but different labels. These are issues familiar to library cataloguing and authority control, and knowledge organization systems such as thesauri.

RDF therefore specifies that the subject and predicate of a triple must be represented by a machine-processable identifier, the URI. The quality required of a URI is uniqueness. A URI should only refer to one thing to avoid the ambiguity problem, although one thing may be identified by many URIs. And a URI should not be changed within a

triple after it has been used, to maintain consistency and coherence. A triple gains utility when it is linked to other triples with the same subject in a set of statements about the same thing; the integrity of the set is destroyed if any of the URIs is amended. A URI should be persistent once it is used. Any unique string of machine-processable characters can be used as a URI, but extra functionality can be obtained if the string uses the same syntax as the uniform resource location (URL) used to identify documents on the web. A URI in this form can be processed as if it were a URL, using existing infrastructure, to obtain either a machine-readable or human-readable version of a document giving more information about the thing identified. This technique is known as "de-referencing" the URI.

The object of a triple may also be a URI, but literal strings such as labels and text paragraphs are allowed, as well as typed strings such as numbers and dates. Again, more utility is gained if the object is a URI, so that the whole triple is composed of three URIs. This allows a machine to not only gather sets of triples with the same subject, but also chains of triples where the object of the first triple is the same as the subject of the next, and so on. Of course, use of a URI for the object also resolves the ambiguity problem. However, there are many situations where the object is unlikely to be the subject of another triple, for example a statement of responsibility in a book, so the overhead of assigning a URI and creating an extra triple which associates it with the literal text can be avoided. If there is a need to assign a URI subsequently, an additional triple using the URI can be added. For example, our "this paper"-"author"-"Gordon Dunsire" triple may be represented for RDF as S1-P2-"Gordon Dunsire", with the subject and predicate represented by the URIs S1 and P2 respectively (these are not examples of real URIs). Later on, the thing or entity with the label "Gordon Dunsire" might be given the URI S8. It is not good practice to change the original triple, but it is acceptable to create a new triple S1-P2-S8. A machine can easily decide which triple is more useful if its task is to create triple chains.

RDF can also represent how the properties used as predicates relate to one another, for example that "creator" is a more general role than "author", as well as what categories of thing a property is restricted to. Categories which define a commonality between different things are represented by classes. Any thing can belong to many classes: the class of male humans, the class of humans born in Scotland, the class of family names beginning with "D", etc. In RDF, all things are a member of the class Thing (classes are conventionally named with an initial uppercase letter, while property names start with a lowercase

letter). RDF Schema [4] and Web Ontology Language (OWL) [5] are the main tools for constructing relationships among and between classes and properties to create "knowledge structures"; both are applications of RDF itself, so everything, the content of knowledge and the structure which carries it, can be represented as triples.

A triple is essentially metadata, as indicated by the use of the term "subject" for the first part. The statement represented by a triple is "about" the subject; it is data about data.

2. Library expertise

The professional practice of librarianship, in particular cataloguing, classification, and indexing, has well over 150 years of experience and expertise in developing metadata schemas and creating structured metadata, if the 1841 date of Panizzi's rules for the catalogue of the library of the British Museum is taken as the beginning of the modern approach. This approach strives to be logical and structured while fulfilling the primary goal of supporting bibliographic information retrieval services maintained and used by humans. As practice has evolved during the twentieth century, there has been an increasing consensus on the basic entities, attributes, and relationships of metadata required for retrieval services; but there has also been a proliferation of different record formats, controlled terminologies, and knowledge organization systems. Much of the ecology of bibliographic metadata has been driven by the rapid and far-reaching impact of technology on the information environment, starting with the introduction of machine-readable cataloguing (MARC) in the 1960s [6]. As might be expected, new formats and systems have appeared, to fill new niches that well-established implementations have been unable or slow to adapt to. Adaptation has, nonetheless, resulted in some consolidation of earlier differences, as exemplified by the evolution of MARC from the original USMARC format to the current MARC21, and the extinction of UKMARC and other variations.

The increasing availability in the past 50 years of technology for capturing, storing, processing, and exchanging bibliographic metadata by computer has resulted in large quantities of machine-processable records. It is very difficult to estimate how many digital catalogue records exist. The largest aggregation of records is OCLC's WorldCat [7], which contains over 200 million records, most of which are in MARC21 format. It is probably safe to assume that there are one billion records in digital form or a physical form which can be digitized with

little human intervention, such as printed or typed inventories of manuscripts and archives. There is significant duplication in this metadata: the dialectic between local and centralised data storage results in central records being copied to local systems with only partial subsequent recombination into a consortial master record. The extent of duplication is, again, very difficult to estimate.

Libraries have therefore amassed large numbers of high-quality metadata records pertaining to bibliographic resources, including online digital resources as well as the physical manifestations of printed and manuscript works. The number of metadata elements in a record depends on the format and rules for creating content, varying from five to fifty, corresponding to the lowest and highest of the levels suggested by the Anglo-American cataloguing rules, second edition (AACR2). They are selected from schemas containing between 15 (Dublin Core) and 250 (MARC) distinct elements. With a very rough assumption that an average record will contain 10 elements and each element can be represented by one triple, library legacy records could yield 10 billion triples for the Semantic Web. This is probably an underestimate.

Furthermore, a high degree of linkage can be established between these triples, through the use of existing and potential mappings between metadata elements in different schemas and formats, and direct comparison of identifiers such as ISBN that are embedded in records. Automated statistical correlation is also a useful tool when such large numbers of data are available, as demonstrated by OCLC's experimental Classify service [8] and the mappings between Dewey decimal classification (DDC) notations and Library of Congress subject headings (LCSH) available in OCLC's WebDewey service [9].

Library metadata expressed in RDF may therefore provide a critical mass of linked data useful for the development and utility of the Semantic Web. In particular, library knowledge organization and authority control systems can supply rich and comprehensive structure and content to aid the identification of specific entities of general interest to Semantic Web applications, including persons, organizations, places, and intellectual concepts. Examples of controlled terminologies already represented in RDF, and therefore with assigned URIs, include the Virtual International Authority File (VIAF) for personal names. and subject headings in English (LCSH) [10], French (RAMEAU) [11], and German (SWD) [12]. The known issue of duplication may also have some benefit in helping the research and development of algorithms for detecting and avoiding the processing of sets of triples representing the same semantic information; library systems developers have been struggling for years to find ways of

matching metadata in record aggregations such as union catalogues, and the nature of the problem is well understood.

3. Library standards

Librarianship has undertaken recent analyses, in the last 10 years or so, of the metadata requirements to support user-centred knowledge discovery services in the digital environment. The first important outcome was Functional requirements for bibliographic records (FRBR) [13], originally published in 1998. This identifies the basic tasks undertaken by users of a bibliographic information retrieval service and provides a model of the entities, attributes, and relationships required to meet them. Supplementary models for authority records [14] and subject authority records [15] have been published subsequently. RDA: resource description and access [16], is a new set of guidelines and instructions on formulating data to support resource discovery, covering the full range of resource content and media. It is a successor to AACR2, and is based on the FRBR model.

The International Standard Bibliographic Description (ISBD) has also undergone extensive analysis and development, culminating in the draft consolidated edition [17]. ISBD was originated in 1969 and has influenced many library metadata standards since, including AACR2 and MARC; it continues to be directly related to the UNIMARC format. The consolidated edition is also aligned with FRBR where appropriate.

Thus there is a set of modern, interconnected models and their applications for library metadata. It is by no means completely coherent because of the independent evolution of the components [18], but there appear to be no insurmountable barriers to improving this.

These metadata models and structures are in the process of being represented in RDF [19, 20]. In general, an entity such as FRBR's Work is treated as a class, an entity attribute such as ISBD's common title is treated as a property with the entity class as domain and nothing specified as the range, and an entity relationship such as FRBR's supplement as a property with the related entity classes specified as domain and range respectively. A property's domain and range constrain the semantic inferences and links that can be made between two or more triples. Additional constraints are being represented in RDF as an OWL ontology or a Dublin Core application profile [21]. The connections between the different models can also be represented as triples with OWL properties such as equivalentClass and equivalentProperty. The same method can used to connect these

library structures to related RDF classes and properties used by other communities, such as FOAF (friend of a friend) [22] and Simple Knowledge Organization System (SKOS) [23]. These are often broader in definition, and will provide a bridge between library metadata and the rest of the Semantic Web.

The RDF representation of specific models should help the development of representations of the formats in which legacy library metadata is currently stored; for example, the work on ISBD is expected to inform the development of an RDF version of UNIMARC. And a representation of the format will help the parsing of legacy records into triples.

4. Current activity

Several organization have initiatives underway to produce triples from legacy records, most if not all in MARC21 format. These include the British Library [24], Mannheim University [25], and the National Library of Sweden [26]. In most cases only a subset of the record format's attributes have been mapped to RDF properties, and most mappings are lossy, for example when all subfields of a MARC tag collapsed to a single property. Mappings have been constrained by the limited availability of suitable RDF properties; both the British Library and Mannheim University projects are using ISBD properties which are have not yet been officially approved by the ISBD Review Group. Properties have been selected from a variety of sources, including Bibliographic ontology, Dublin Core, and Dublin Core terms.

The W3C initiated a Library Linked Data Incubator Group [27] in June 2010 with the mission "to help increase global interoperability of library data on the Web". It envisions libraries as "a potential major provider of authoritative datasets (persons, topics...) for the Linked Data Web". The Group is expected to produce a number of deliverables in May 2011, including a use case document based on real-world activity, requirements for improving integration of library environments with the Semantic Web, and a snapshot of current technologies, vocabularies, and ontologies relevant to library linked data to help identify what extensions and other standards may be needed.

The Group has discussed many of the topics raised in this paper, and is in the process of identifying associated problems and solutions. These include the assignation of URIs to legacy metadata and record granularity; for example, a record can be given a single URI as a member of the ISBD class Resource and mapped to ISBD properties,

but must be split into a four parts if it is to use any FRBR properties, which have one of the FRBR classes Work, Expression, Manifestation, or Item has domain. Each of these parts must be assigned a separate URI; the URI for the whole record cannot apply to any separate part. A related problem is the duplication of legacy records, which will be naively assigned different URIs as they are recast as triples. This will result in large numbers of triples which appear to have different subjects but are actually the same, resolved if the duplication can be avoided in the first place or equivalence properties linking the URIs are created. Another issue is interoperability and mapping between the classes, properties, and vocabularies in RDF representations of different library models and standards, and between library representations and those of other communities.

5. Impact

The engagement of libraries and librarianship with the Semantic Web, if fully realised, will have a significant impact on current practice and services. The adoption of information technology in the 1960s sharpened the focus of metadata management on the catalogue record by removing the need to create physically separate entry points to retrieval indexes, such as catalogue cards. Although rules for determining metadata content remained focussed on parts of the record such as description and headings, for example AACR2, the systems and workflows for metadata creation and maintenance were, and are, designed for the record as the unit of processing. Authority control using separate records for headings is largely maintained at national rather than local level; if carried out in a local library, it is usually as an on-the-fly operation embedded in bibliographic record maintenance procedures. In the Semantic Web, the focus is on the triple. Triples are designed to stand alone as complete statements, so there is no particular need to store or process sets of triples as a unit. Of course, a set of triples with the same subject is required for display of a record-scale view of the metadata; this is the main reason for the FRBR model of four component "records", as each can be displayed separately during specific phases of a search to support relevant user tasks. But there is no requirement that all triples in a set are derived from the same source or domain. Indeed, a major benefit to libraries in the Semantic Web is the reusability of metadata from outside of the library environment, such as the publishing and bookselling communities and online reference works like dictionaries and encyclopaedias. A shift of focus from the record to the triple will require re-engineering most of the current library metadata management

support infrastructure, including software, data storage and processing, workflows, and interaction with other communities, as well as user interfaces and information retrieval services.

Another impact, perhaps more unsettling for cataloguers, is likely to be the incorporation of metadata from non-professional sources, specifically end-users and machines. Social networking services are generating large quantities of user-generated data about online bibliographic resources, including book reviews, subject terms attached to photographs and videos, lists of contents of music CDs, film genres and categories, etc. These metadata are often subjective, ill-informed, or otherwise unreliable, but if the quantity is large enough a regression to the mean can be expected: a consensus will emerge by applying statistical processes, which may be further strengthened by linking vernacular terms to controlled vocabularies. This approach has worked well for search engine optimisation in the World-Wide Web, and is the basis for the library-oriented initiatives mentioned previously. The law of large numbers should raise the quality of user-generated metadata to the same level as other non-library sources such as publishers, sufficient to be useful in library services.

Metadata created automatically by machines from digital resources is becoming an important adjunct to the practice of professional cataloguing. Much of the content of descriptive bibliographic metadata is obtained by transcribing information from the resource itself, in accordance with standard cataloguing rules such as AACR2. The exactness of transcription is traditionally diluted by issues such as shortening lengthy text, interpolating explanatory text, treatment of abbreviations and special characters, and, in the case of Latin alphabet scripts, letter case: text all in upper case, capital letters is generally unacceptable to users. Machines are much more efficient and effective than humans in transcribing digital or digitized text exactly; it is not so much transcription as duplication. This has been reflected in RDA which encourages the use of "as you see it" content for appropriate FRBR attributes. Machines can assist when an option is taken in order to improve user-friendliness, such as recasting upper case text into so-called "title" case, where computers can assist by identifying acronyms and other groups of letters which should remain in upper case, and otherwise applying a simple algorithm to change the case of the rest.

A powerful feature of RDF is the ability to infer other triples from a specific triple. For example, if P1 is a property with domain and range specified as classes C1 and C2, then a triple using this property, say A1-P1-B1, can be used to infer the triples A1-"is a member of the class"-C1 and B1-"is a member of the class-"C2". The inferred triples

do not need to be explicitly declared. Software for semantic processing of triples can use such inferred triples to fill gaps and detect anomalies. If the classes C3 and C4 have been declared disjoint in an OWL ontology, and inferred triples from two different properties show that a URI is a member of both C3 and C4, then there is a logical inconsistency in the source triples. Machine-generated metadata will also be useful for library services, and cataloguers need to be familiar with its characteristics.

Library linked data has significant potential benefit for users of the general Web (including Web 2.0 and the Semantic Web) and of bibliographic information retrieval services, especially if the data is "open" for non-commercial use. The availability of library knowledge organization systems for in-depth, specialised subject vocabularies, as well as those with a broad coverage like LCSH and DDC, will improve the Semantic Web's capability for forming clumps of semantic cohesion. The existing semantic relationships in legacy records and the infrastructure for maintaining them between distributed metadata statements, in the form of new triples, may provide a much richer context for the current keyword-based search engines for Web documents. Methods for partial semantic mark-up of document content, such as RDFa [28], improve the context of specific entities in relation to their occurrence in other documents. Bibliographic properties specify a comprehensive set of relationships between documents themselves, improving cohesion at the granularity level of sets or collections of documents. Other properties specify relationships between documents and agents or parties who are also related to document content at the semantic and non-semantic keyword levels, thus bridging different layers of granularity.

Information retrieval services offered by libraries will benefit from metadata from non-bibliographic contexts. Links to book covers, reviews, and bookshop stock data are becoming commonplace in modern public access catalogues. Mashups of bibliographic data and geographical location data are becoming familiar to users; examples include WorldCat and the Scotland's Information service [29]. Services in a linked data Semantic Web environment will benefit from similar conflations of metadata, such as biographical data about an author, encyclopaedic data about a subject, or reviews of a title, but in much more flexible ways with much larger amounts of data.

6. Conclusion

Libraries have a significant role to play in the Semantic Web, and the Semantic Web is likely to have a far-reaching impact on librarianship and library services. The relationship is symbiotic; each needs to exchange data, functions, and methods with the other.

References

[1] BERNERS-Lee, Tim, Hendler, James, Lassilla, Ora. (2001). **The Semantic Web**. *Scientific American*, May 17, 2001. Available at http://www.scientificamerican.com/article.cfm?id=the-semantic-web (Accessed on 24/01/2011)

[2] W3C. **Resource description framework (RDF)**. Available at http://www.w3.org/RDF/ (Accessed on 24/01/2011)

[3] W3C. **Uniform resource identifier (URI): Generic syntax**. Available at http://tools.ietf.org/html/rfc3986 (Accessed on 24/01/2011)

[4] Available at http://www.w3.org/TR/rdf-schema/ (Accessed on 24/01/2011)

[5] W3C. **OWL Web Ontology Language: overview.** Available at http://www.w3.org/TR/owl-features/ (Accessed on 24/01/2011)

[6] LIBRARY OF CONGRESS. NETWORK DEVELOPMENT AND MARC STANDARDS OFFICE. **MARC standards**. Available at http://www.loc.gov/marc/ (Accessed on 24/01/2011)

[7] OCLC. **WorldCat window to the world's libraries**. Available at http://www.oclc.org/uk/en/worldcat/default.htm (Accessed on 24/01/2011)

[8] OCLC. **Classify: an experimental classification web service**. Available at http://classify.oclc.org/classify2/ (Accessed on 24/01/2011)

[9] OCLC. **WebDewey: the easiest way to use DDC**. Available at http://www.oclc.org/dewey/versions/webdewey/ (Accessed on 24/01/2011)

[10] LIBRARY OF CONGRESS. **Library of Congress subject headings**. Available at http://id.loc.gov/authorities/ (Accessed on 24/01/2011)

[11] STITCH. **RAMEAU subject headings as SKOS linked data**. Available at http://www.cs.vu.nl/STITCH/rameau/ (Accessed on 24/01/2011)

[12] DEUTSCHE NATIONALBIBLIOTHEK. (2011). **The linked data service of the German National Library.** Version 3.0. Available at http://files.d-nb.de/pdf/linked_data_e.pdf (Accessed on 24/01/2011)

[13] IFLA STUDY GROUP ON THE FUNCTIONAL REQUIREMENTS FOR BIBLIOGRAPHIC RECORDS. (1998). **Functional requirements for bibliographic records**. Amended 2009. Available at http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records (Accessed on 24/01/2011)

[14] IFLA WORKING GROUP ON FUNCTIONAL REQUIREMENTS AND NUMBERING OF AUTHORITY RECORDS (FRANAR). (2009). **Functional requirements for authority data**. Available at http://www.ifla.org/publications/functional-requirements-for-authority-data (Accessed on 24/01/2011)

[15] IFLA WORKING GROUP ON THE FUNCTIONAL REQUIREMENTS FOR SUBJECT AUTHORITY RECORDS (FRSAR). (2010). **Functional requirements for subject authority data (FRSAD): a conceptual model**. Available at http://www.ifla.org/files/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf (Accessed on 24/01/2011)

[16] **RDA toolkit**. Available at http://www.rdatoolkit.org/ (Accessed on 24/01/2011)

[17] ISBD REVIEW GROUP. (2010). **International standard bibliographic description (ISBD)**. Consolidated edition. Draft as of 2010-05-10. Available at http://www.ifla.org/files/cataloguing/isbd/isbd_wwr_20100510_clean.pdf (Accessed on 24/01/2011)

[18] DUNSIRE, Gordon. (2009). **UNIMARC, RDA and the Semantic Web**. Presented at World Library and Information Congress: 75th IFLA General Conference and Assembly, 23-27 August 2009, Milan, Italy. Available at http://www.ifla.org/files/hq/papers/ifla75/135-dunsire-en.pdf (Accessed on 24/01/2011)

[19] DUNSIRE, Gordon, Willer, Mirna. (2010). **Initiatives to make standard library metadata models and structures available to the Semantic Web.** Presented at World Library and Information Congress: 76th IFLA General Conference and Assembly, 10-15 August 2010, Gothenburg, Sweden. Available at http://www.ifla.org/files/hq/papers/ifla76/149-dunsire-en.pdf (Accessed on 24/01/2011)

[20] WILLER, Mirna, Dunsire, Gordon, Bosančić, Boris. (2010). **ISBD and the Semantic Web**. *JLIS.it.,* vol. 1, no. 2 (Dicembre/December 2010), 213-236. Available at http://leo.cilea.it/index.php/jlis/article/view/4536/4408 (Accessed on 24/01/2011)

[21] COYLE, Karen, Baker, Thomas. (2009). **Guidelines for Dublin Core application profiles**. 2009. Available at http://dublincore.org/documents/profile-guidelines/index.shtml

[22] FOAF PROJECT. **The Friend of a Friend (FOAF) project**. Available at http://www.foaf-project.org/ (Accessed on 24/01/2011)

[23] W3C. **SKOS simple knowledge organization system - home page**. Available at http://www.w3.org/2004/02/skos/ (Accessed on 24/01/2011)

[24] BRITISH LIBRARY. METADATA SERVICES. **Sample data**. Available at http://www.bl.uk/bibliographic/datasamples.html (Accessed on 24/01/2011)

[25] UNIVERSITÄTSBIBLIOTHEK MANNHEIM. **Linked data service (public beta)**. Available at http://data.bib.uni-mannheim.de/index_en.html (Accessed on 24/01/2011)

[26] MALMSTEN, Martin. (2008). **Making a library catalogue part of the Semantic Web**. *2008 Proceedings of the International Conference on Dublin Core and Metadata Applications*. Available at http://dcpapers.dublincore.org/ojs/pubs/article/download/927/923 (Accessed on 24/01/2011)

[27] W3C LIBRARY LINKED DATA INCUBATOR GROUP. Available at http://www.w3.org/2005/Incubator/lld/Overview.html (Accessed on 24/01/2011)

[28] W3C. (2008). **RDFa primer: bridging the human and data webs**. Available at http://www.w3.org/TR/xhtml-rdfa-primer/ (Accessed on 24/01/2011)

[29] **Scotland's information**. Available at http://www.scotlandsinformation.com/ (Accessed on 24/01/2011)