Article can be accessed online at http://www.publishingindia.com

BUILDING A FEDERATED DIGITAL LIBRARY: APPROACHES TO METADATA INTEGRATION

Richard Gartner*

Abstract *Two approaches to integrating metadata in federated digital environments are discussed, one based on semantic web methodologies, the other on XML architectures. It is argued that it is possible to emulate the semantic flexibility of the semantic web within XML-based schemas by using linkages to controlled vocabularies using Universal Resource Identifiers (URIs). The advantages of retaining XML as the basis of digital library metadata remain clear, particularly when semantic linkages of this type can be integrated into XML architectures.*

Keyword: *Metadata, Linked Open Data, XML, Digital Libraries*

INTRODUCTION

Perhaps the most seismic shift undergone by the library community over the last 100 years has been the advent of the 'digital library' which has seen extensive collections converted to machine-readable form and disseminated on scales and with a speed which would have been unthinkable in the traditional 'analogue' library environment. The expansion of digitization programmes, like that of the universe itself, shows no sign of slowing down: a survey, for instance, by the US-based Association of College and Research Libraries concluded in 2010 that the "digitization of unique library collections will increase and require a larger share of resources" requiring that "libraries often must reallocate fiscal resources to support these projects" (ACRL Research Planning and Review Committee, 2013).

The extensive resources now flowing into digitization projects have changed completely the notion of where the boundaries of a library now lie. The same report rightly concludes that "the definition of the library will change as physical space is re-purposed and virtual space expands" as federated environments, in which the resources of potentially many administratively and physically dispersed collections, become findable for the first time as single, coherent resources. Moving from an individual digital library to a federated environment, in which digital objects can be searched for and accessed wherever they are located, does, however, require an integrated approach to metadata.

LIBRARY METADATA: ANALOGUE AND DIGITAL

In the traditional library, metadata integration, and the consequent interoperability between records, advanced greatly in the mid-1960s when the MARC standard was devised to facilitate the computer-readable cataloguing of library materials. Standardising a machine-readable format for the interchange of catalogue records completely transformed the relationship of libraries to each other: for the first time, the sharing of records allowed libraries to co-operate in everything from the creation of cataloguing information to the federation of resource discovery. Without this standardisation of cataloguing practice, many key features of the modern bibliographic environment, including large-scale union catalogues such as Worldcat ('WorldCat.org: The World's Largest Library Catalog', n.d.) and co-operative cataloguing ventures such as OCLC ('OCLC: Worldwide, member-owned library cooperative | United Kingdom and Ireland', n.d.) would have proved impossible.

Unfortunately, such metadata interoperability has so far proved elusive in the digital library world. Part of the reason for this is the greater complexity of digital library metadata which makes the provision of a single metadata standard, analogous to MARC, impossible to achieve. The bulk of a MARC record comprises bibliographic information necessary to find a book on the shelves, supplemented by a small amount of internal administrative information (such as accession numbers). This limited element set, however, rapidly proves inadequate for handling complex digital objects.

The National Information Standards Organisation defines three types of metadata required to support a digital object: descriptive metadata (used for resource discovery and identification), structural metadata (used to to show how compound objects are put together by expressing the relationships between their component parts), and administrative metadata (used to administer the object, including rights management and preservation information) (National Information Standards Organization, 2004, p.

* Centre for e-Research, King's College, London, United Kingdom

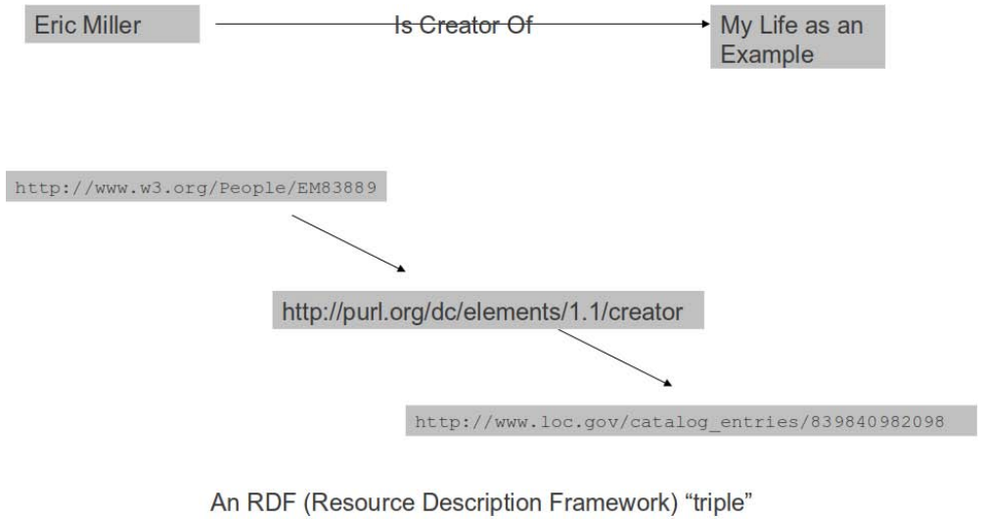

Figure 1: Encoding a semantic relationship in RDF

1). All three types are usually needed in a digital library environment: in addition to the descriptive metadata in the standard MARC record, digital library metadata must also provide complex technical, rights and other administrative information, and include structural metadata to allow the internal navigation of complex objects when they are delivered to the user.

Because of these more complex requirements, a single metadata scheme, analogous to MARC, is likely to prove insufficient to meet them all. An integrated environment therefore almost certainly requires the use of multiple schemes used in conjunction: to allow these to work together it is essential to utilise a mechanism for establishing linkages between their internal components. Such an approach must also work at multiple levels of granularity which extend beyond the metadata requirements of a single item. To allow federated searching and browsing in particular, not only must a consistent metadata strategy be implemented within the digital collection but metadata must be applied which is at least inter-operable between collections in the wider environment within which they are located. Even in a world in which such services as Google can offer a full-text approach to finding disparately located materials, it is important for collections to apply consistent and established metadata standards to allow their collections to inter-operate in this way.

APPROACHES TO INTEGRATION

Two approaches may be taken to integrating this diverse metadata environment, each of which takes a diametrically opposed view to building up an integrated architecture. One method, based on the methodologies constructed for the semantic web, is essentially atomistic in its overall design: metadata components at the lowest level of granularity are linked to others in flexible architectures based on semantic linkages, usually employing RDF (Resource Description Framework) 'triples' to encode these.

An RDF triple is a set of three elements representing the subject, predicate and object of a semantic assertion: each of these is often represented by a Universal Resource Identifier (URI), which provides a precise identification of their meaning. In this example, three URIs are used to identifier a person as the creator (as defined by Dublin Core) of a work.

```
<employee>
    <name>Eric Miller</name>
    <employee_number>291919</employee_number>
    <email_address>e_miller@nowhere.ac.uk</email_address>
    <personal_title>Dr.</personal_title>
    <publications>
    <publication>My Life as an Example</publication>
    <publication>My Life as an Example: volume 2</publication>
    </publications>
</employee>
```

Example 1: encoding information in an XML hierarchy

These triples may themselves be linked to others as subjects or objects of another triple. In this way, they form the

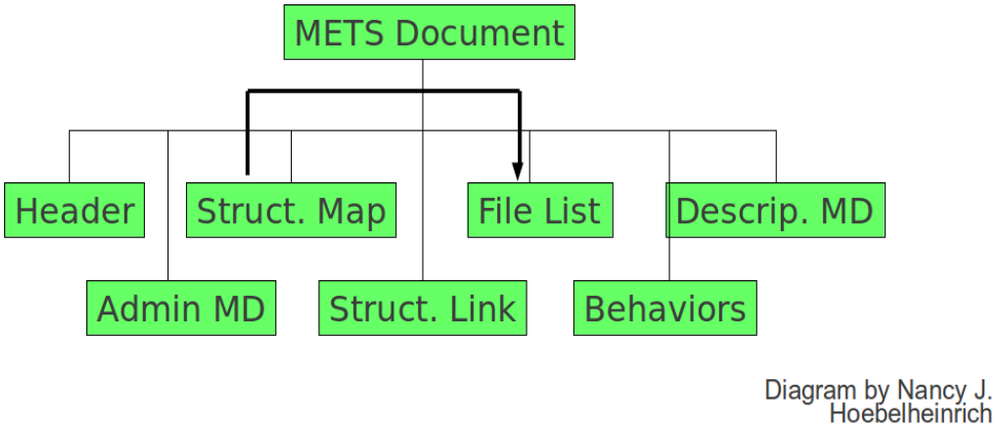

Figure 2: Internal structure of a METS file

atoms from which an entire metadata architecture may be constructed, linking together all of its components into a semantically-connected whole. This method underlies the popular Fedora Commons digital repository system ('Home — Fedora Repository', n.d.) and is also used by the Library of Congress to publish their key standards and vocabularies as Linked Open Data ('Home - LC Linked Data Service (Library of Congress)', n.d.).

The alternative approach, which predates the semantic web and RDF, uses the XML syntax to "package" metadata in prescribed hierarchical architectures. A set of metadata on the same person may be encoded in XML as follows:-

In this case, a strict hierarchy is used to encode the relationships between components of the metadata record, instead of the network of semantic linkages offered by RDF. The hierarchical foundations of XML, which are owed to its origins as a language for encoding texts, appear inflexible compared to more fluid relationships which may be encoded in RDF: they may be tempered, however, by the use of an extensive set of linking mechanisms, including hyperlinks, which the XML syntax is capable of encoding.

The XML "packaging" methodology replies on the ability of XML to allow documents conforming to multiple schemas to be co-located within a single document: this is done by the use of "namespaces", a method for prefixing each element with a URI (or a shortened name in place of the URI) to delineate the schema to which it belongs. This methodology underlies the widely-used Metadata Encoding and Transmission Standard (METS) ('Metadata Encoding and Transmission Standard (METS) Official Web Site', n.d.) packaging standard for digital library metadata, which uses XML architectures to link together metadata held in disparate XML schemas within a single framework.

METS uses the namespace mechanism to allow metadata encoded in any XML schema to be embedded directly in its architecture within one of top-level sections which divide metadata according to type (one for descriptive metadata, one for administrative etc). XML linkages are then used to express relationships between these metadata components: in this diagram, for instance, such a link may join a component in the METS structural map (which expresses the internal structure of a complex object) to an entry in the file list indicating the location of the file (for instance an image of a page of a book) corresponding to this component.

METS, and its associated XML-based methodology, is now well established in digital libraries and is extensively used in some major implementations (Library of Congress, 2011a). Nonetheless, some criticisms have been levelled about its hierarchical rigidity which may limit the flexibility of users' search options (Han, 2006, p. 236). In a federated search environment, inflexibilities of this type could prove major impediments to the effectiveness of an integrated digital library. For these reasons, the use of semantic web methodologies is the subject of much research in the library community at present, and has even been called a "beckoning paradise" by some in the digital preservation community (Gonzalez, 2011).

Nonetheless, despite several years of development and intensive research, the implementation of semantic web approaches as the basis for metadata in working systems still proves relatively elusive, despite its adoption by such systems as Fedora Commons. Much of the reason underlying this is the complexity of handling the atomistic metadata objects that form the basis of the semantic web, which may amount to several thousand RDF "triples" even to describe a single digital object. This may prove particularly problematic when interchanging or migrating records between systems: in such cases, transferring a network of small components is much more complicated than a single, packaged XML file. Research, for instance, has found that the claimed theoretical interoperability of RDF-based Fedora Content Models (FCM) has proved difficult to achieve in practice (Sharma, 2007).

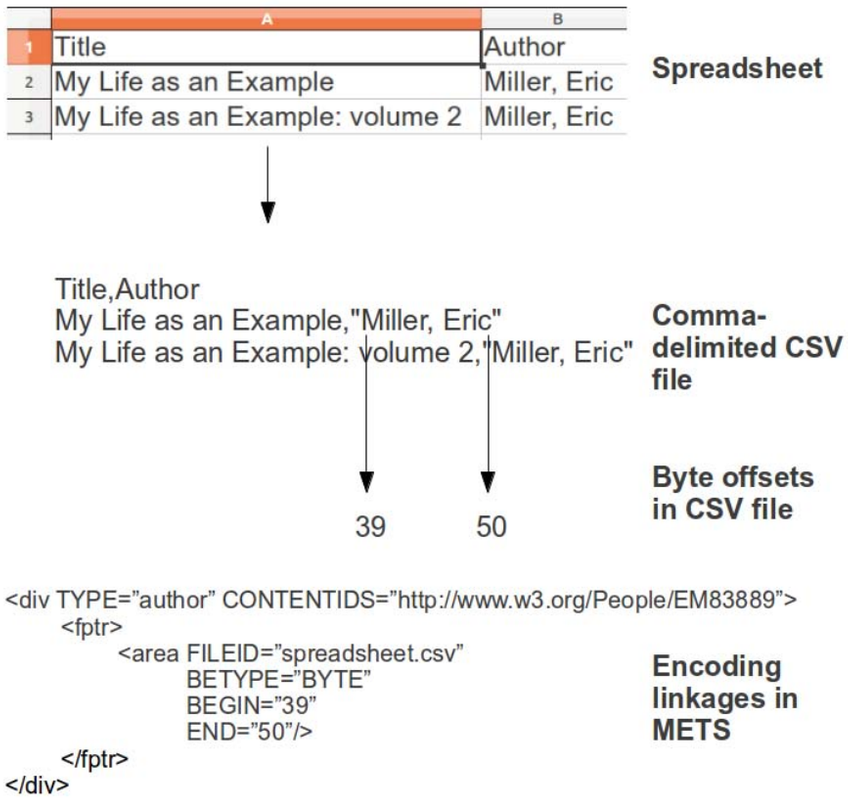

Figure 3: Encoding semantic relationships in METS

ACHIEVING SEMANTIC INTEROPERABILITY WITH XML

Despite the relative ease-of-use that packaged XML may offer over RDF, some work needs to be done to allow it to duplicate some of the flexible functionality that the semantic web offers. In particular, the ability to record semantic as opposed to syntactic relationships is often cited as a key advantage of the latter approach (Patel-Schneider & Siméon, 2002), particularly to allow federated semantic searching across disparately-located resources.

Despite the undoubted ease with which RDF may be used to enable this functionality, it is possible to emulate it within XML architectures and so retain the advantage of the flexibility offered by RDF while retaining the ease-of-use of packaged XML. To do so requires the use of XML's extensive linking mechanisms which allow components of an XML file to be identified and linked to at any level of granularity.

Semantic relationships may be easily encoded using XML-based authority lists or controlled vocabularies. To do so requires the use of a schema constructed for encoding such vocabularies, such as MADS (Metadata Authority Description Schema) (Library of Congress, 2011b), which is maintained, like METS, by the US Library of Congress. MADS allows thesauri and other vocabularies to be encoded within logical architectures, and, crucially, allows a Universal Resource Identifier (URI) to be assigned to every term. These URIs allow the same concept to be identified wherever it appears in the global metadata environment. A sample thesaurus entry for the fictional author in Example 1 may take this form:-

The **valueURI** attribute for the <name> element defines the URI which may be used to identify this person.

To provide semantic linkages within the XML environment it is then necessary to use a schema designed to allow referencing of this type. Ideally, such linking should be possible at any level of granularity if this methodology is to achieve the same level of flexibility as the semantic web and RDF allow. Fortunately, the METS schema described above does fulfill both of these criteria and can be used with great flexibility to allow semantic linkages of this type.

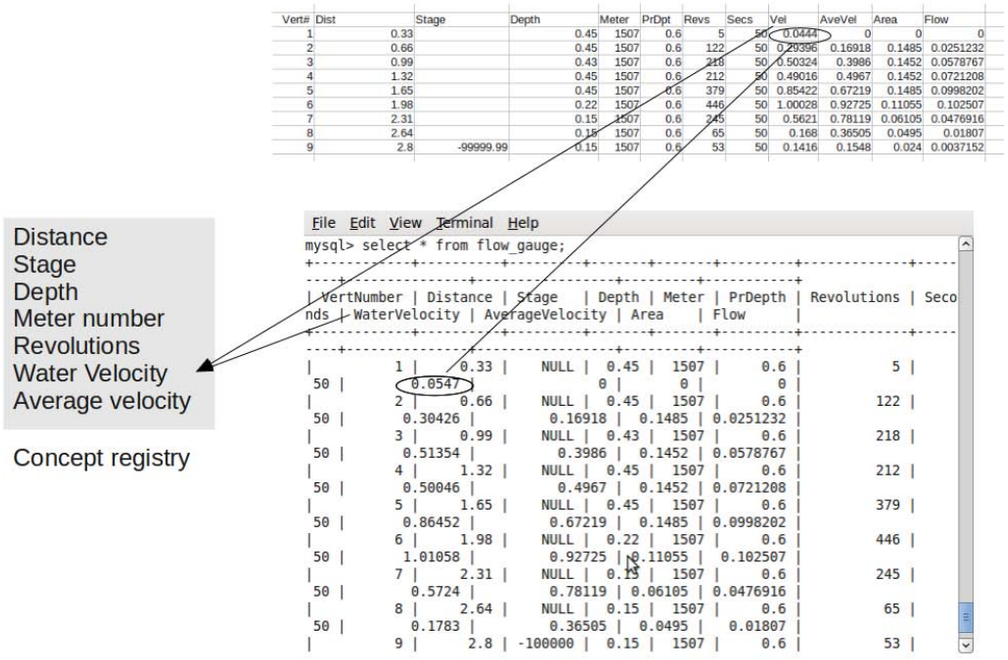

Figure 4: Mapping of disparately labelled concepts

```
<mads>
  <authority>
   <name type=”personal”
   valueURI=”http://www.w3.org/People/EM83889”>
   <namePart type=”family”>Miller</namePart>
        <namePart type=”given”>Eric</namePart>
     <namePart type=”termsOfAddress”>Dr</namePart>
     </name>
  </authority>
</mads>
```

Example 2: name authority entry encoded in MADS

METS allows the internal components of any file that makes up a complex digital object to be referenced individually at any level of granularity. Any component of the structural map, the core of a METS file which describes the internal structure of an object and the relationships between the files that make it up, can be subdivided further by an element named <area>: these components may include, for instance, a part of a textual object, a data item in a database or spreadsheet, a time-coded section of a video file or a part of an image. METS then allows these components to be associated with a URI containing a controlled term for their content: for instance, the label in a column in a spreadsheet may be associated with a URI for a term describing the semantics of the content of the column.

The diagram above demonstrates how these links may be made at a fine level of granularity. In a spreadsheet, the works of our fictional author are listed as separate rows: the spreadsheet is then saved as a comma-delimited CSV (Comma-Separated Values) file, a simple text file in which each column in the spreadsheet is separated from the others using commas. The beginning and end of each entry can be specified by noting the byte offset (the number of bytes from the beginning of the file) where each starts (in this case, the first entry for the author’s name begins at byte offset 39 and ends at 50).

In the METS file, the author’s name in the spreadsheet entry is encoded by listing the start and end byte offsets for the entry in the BEGIN and END attributes within the <area> element. This is, in turn, nested within a (division) element which co-locates all entries corresponding to this name: the semantic linkage to the controlled vocabulary is done by noting the URI for this person in the CONTENTIDS attribute. In this way, the entry “Eric Miller” in the spreadsheet is semantically identified as corresponding to the “Eric Miller” in the thesaurus controlled vocabulary (where he is identified by the valueURI attribute).

Forming linkages in this way, between data components at low levels of granularity and controlled terms in XML-encoded vocabularies, it becomes possible to emulate the semantic interoperability of RDF within a packaged XML environment. The set of linkages shown in this example

equate semantically to the RDF triple shown in Figure 1.

Although this approach is less elegant, and certainly more verbose, than an RDF 'triple', it offers considerable advantages once objects achieve any significant level of complexity. The hundreds or thousands of triples needed to encode the metadata for a compound digital object rapidly become cumbersome to manage and to transfer between systems. The METS approach allows all of the metadata for a highly complex object to be packaged neatly into a single file without losing any of the semantic richness which its RDF counterpart allows.

This approach has already been tested in the context of digital archives and found a viable methodology for handling the metadata requirements of complex and diverse data. One example is the UK-based Demonstration Test Catchments Archive (National Demonstration Test Catchment Network, 2011) of environmental data, which aims to build an archive of low-level data related to water quality in British rivers. The data is collected from a wide range of discrete sources in multiple formats and is not catalogued in any coherent fashion. To allow interoperability between these diverse sources requires the same concept to be semantically identified wherever it occurs in the data. Figure 4 illustrates the problem: the same concept (in this case 'water velocity' must be mapped to a central concept registry whether it is called 'Vel' (in the spreadsheet at the top) or 'WaterVelocity' (in the database at the bottom).

This may readily be achieved by the use of METS which record the byte offsets of each component of the CSV files generated from these spreadsheets and databases, and map these to a central 'concept registry' encoded in MADS. All of these mappings may be packaged within a single METS file, which is readily transferred between systems and is archivally robust for preservation purposes.

If required, such an approach may also be used to generate RDF triples for use in systems such as Fedora Commons which employ architectures constructed from these. Both approaches can therefore be used in tandem: it is, however, better practice, for the reasons discussed above, to maintain the tightly-structured XML architecture as the primary format for a digital library's metadata and to generate RDF from this as required. This can be compared, albeit somewhat loosely, to the concept of 'entropy' in physics: it is easier to move from a low-entropy (highly structured) environment to a higher-entropy equivalent in which the same information is present but in a highly atomistic state. Although the higher-level structures inherent in a body of RDF-triples can be reconstructed from these, the processes are more complex than those for generating the triples from their more structured counterparts. For these reasons, using XML is likely to prove the less complex and more manageable option for most digital libraries.

CONCLUSIONS

The methodologies outlined in this article offer great potential for realising the potential of federated digital libraries. Using XML architectures to encode linkages in this way allows the easy incorporation of sophisticated semantic structures to be incorporated into the design of digital collections while retaining the rigorous and robust architectures of XML; the result of this is to combine flexibility with ease of maintenance, and minimal demands on development time and personnel. The boundaries between collections rapidly fade as these semantic linkages are established.

Few can deny that the semantic web and its associated RDF mechanisms for data and metadata encoding offer great potential to the librarian, but attention must be paid to the need for strong and tightly-integrated metadata architectures if semantic linking is going to become operational in a federated online environment. The XML syntax remains a viable option for building architectures of this kind, while retaining the possibility of incorporating flexible approaches to semantic linkages at any level of granularity. The challenge in future years is to design schemas which can combine robust architectures and flexibility, so ensuring that the librarian may enjoy the advantages of both.

REFERENCES

ACRL Research Planning and Review Committee. (2013). 2010 top ten trends in academic libraries. *College and Research Libraries News*, *74*(4). Retrieved from http://crln.acrl.org/content/71/6/286.short

Gonzalez, G. (2011). Linked Open Data: A Beckoning Paradise. Retrieved from http://blogs.loc.gov/digitalpreservation/2011/06/linked-open-data-a-beckoning-paradise/

Han, Y. (2006). A RDF-based digital library system. *Library Hi Tech*, *24*(2), 234–240.

Home — Fedora Repository. (n.d.). Retrieved 5 April 2013, from http://fedora-commons.org/

Home - LC Linked Data Service (Library of Congress). (n.d.). Retrieved from http://id.loc.gov/

Library of Congress. (2011a). METS Implementation Registry. Retrieved from http://www.loc.gov/standards/mets/mets-registry.html

Library of Congress. (2011b). Metadata Authority Description Schema (MADS) - (Library of Congress). Retrieved from http://www.loc.gov/standards/mads/

Metadata Encoding and Transmission Standard (METS) Official Web Site. (n.d.). Retrieved from http://www.loc.gov/standards/mets/

National Demonstration Test Catchment Network. (2011). National Demonstration Test Catchment Network.

Retrieved from http://www.demonstratingcatchment-management.net/

National Information Standards Organization. (2004). Understanding metadata. NISO Press. Retrieved from http://www.niso.org/publications/press/UnderstandingMetadata.pdf

OCLC: Worldwide, member-owned library cooperative | United Kingdom and Ireland. (n.d.). Retrieved from http://www.oclc.org/unitedkingdom/en_us/home.html?redirect=true

Patel-Schneider, P., & Siméon, J. (2002). The Yin/Yang Web: XML Syntax and RDF Semantics. Presented at the WWW2002: the Eleventh World Wide Web Conference. Retrieved from http://www2002.org/CDROM/refereed/231/

Sharma, R. (2007). *Fedora Interoperability Review*. London: Centre for e-Research. Retrieved from http://wwwcache1.kcl.ac.uk/content/1/c6/04/55/46/fedora-report-v1.pdf

WorldCat.org: The World's Largest Library Catalog. (n.d.). Retrieved from http://www.worldcat.org/