Metadata Solutions for the Organization of Online Information

By Alexander May

1. Introduction

On February 2, 2010, a Canadian graduate student discovered Haiti's declaration of independence while flipping through a leather-bound binder at the British National Archives.[1] For over 200 years the document had been assumed lost. That the document was found shortly after the devastating earthquake of January 12 has given hope to many Haitian scholars and intellectuals, providing them with the opportunity to review one of the few remaining primary source documents from this period. Although the eight-page pamphlet is now available for any scholar to review online, the troubling fact remains that it was fortuitous that this piece of cultural heritage was found at all. This is not the first time that seminal documents have gone missing from cultural heritage institutions. In March 2009, the Guardian broke the story that the British Library had mislaid almost 9,000 books, some of which have not been seen for over 50 years. Included in this list are several Renaissance treatises on theology, a medieval text on astronomy and quite a few first editions of nineteenth and twentieth century novels. It should be noted that the library believes none of the items were stolen, but rather misplaced somewhere within its 400 miles of shelves.[2]

Needless to say, the failure to provide appropriate cataloging information about these materials contributed to their loss in the physical world. As we begin to provide online access to our collections it is absolutely essential that libraries learn from these mistakes. After all, the British Library only has to cull its 400 miles of shelves to find its lost items; our collections must be found among the 1 trillion unique URLs[3] and 1.3 billion images.[4] To put it another way: it is simply no longer enough to put library content into HTML web pages and expect the data to suddenly become discoverable. Online library content must now employ a variety of standards to ensure the data contained within it can "talk" to the web, and other data sets, as well as comply with the professional standards employed by various cultural heritage institutions. For future

[1] Cave, Damien. "Haiti's Founding Document Found in London." *The New York Times,* (March 31, 2010): Section: World/Americas.

[2] Dawar, Anil and Maev Keendey. "British Library mislays 9,000 books" *The Guardian,* (March 17, 2009): http://www.guardian.co.uk/uk/2009/mar/17/british-library-books-mein-kampf. (Accessed: 2010-04-22).

[3] Alpert, Jesse and Nissan Hajaj. "Official Google Blog: We Knew the Web was Big…" http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html. (Accessed: 2010-04-11).

[4] Ives, Zachary, G. "Scalable Web Crawling and Basic Transactions." University of Pennsylvania CIS 455/555 Internet Web Systems, April 11, 2010.

projects to succeed at Tisch Library, our information must be discoverable, interoperable and highly structured.

Before turning our attention to what exactly metadata is, how it contributes to networked data, and its role in libraries, it may be helpful to reconsider Haiti's founding document. As indicated above, a digital surrogate now exists online at the British National Archives, making it possible for scholars to access it anywhere in the world. That is, as long as they know specifically where it exists. The problem is that British National Archives failed to provide a web page that both encoded and displayed the information according to any known cultural heritage metadata standard. Consequently, the document has been effectively rendered invisible to the web. An online query of the term: "Haitian declaration of independence"[5] returns several news items about its discovery, a Wikipedia article and a link to a travel site, but fails to provide within the first four pages either a direct link to the British National Archives or the document itself. It has only been 3 months since its discovery, and already the digital version of Haiti's Declaration of Independence is becoming invisible, fading into a sort-of digital oblivion invisible to both search engines and libraries alike.

2. Definitions

So what exactly is metadata? One of the better working definitions for the concept originates in a National Standards Information Organization (NISO) publication entitled Understanding Metadata, which states that metadata is: "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource."[6] According to NISO, metadata not only facilitates discovery, but also ensures future interoperability. That is, describing a resource with metadata allows it to be both human and machine-readable. Another working definition of metadata is provided by the Dublin Core Metadata Initiative (DCMI), which states that metadata is: "data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use, usage and preservation."[7] What makes this definition notable is the implied distinction between the types of metadata used in creating an online resource--descriptive, structural and administrative metadata. These definitions, and the intellectual organizing principles they encapsulate, are far more satisfying than the popular definition that metadata is: "data about data."[8] However, to really understand the gist of metadata, and its importance for online information resources, a few additional definitions and examples are in order.

As a concept, metadata is increasingly central to the emerging networked information environment as envisioned by the World Wide Web Consortium (W3C) and Sir Tim

[5] Query performed using Google on April 11, 2010.

[6] NISO. *Understanding Metadata.* Bethesda, MD.: NISO press, 2004. P. 1.

[7] DCMI. "DCMI Glossary-M" http://dublincore.org/documents/usageguide/glossary.shtml#M (Accessed: 2010-04-22).

[8] Ibid.

Berners-Lee. The W3C is the standards organization for the internet, and it frequently uses the term “linked-data” to describe how metadata can better inform the World Wide Web. For the purposes of this paper, therefore, a working definition of the term "linked-data" is necessary. Linked-data may be thought of as metadata that uses the encoding schemes and protocols of the web to both identify information and disambiguate it from like things.[9] According to the W3C, the creation of linked-data is an essential component for future web development.[10] By combining the model of linked-data with the two above definitions for metadata, we can reasonably put forth the idea that while not all metadata is linked-data, *all linked-data is metadata*.

Finally, the additional idea of interoperability is key to both metadata and the related concept of linked-data. Once again NISO provides a nice explanation, stating that interoperability is: “the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality….Using defined metadata schemes, shared transfer protocols, and crosswalks between schemes, resources across the network can be searched more seamlessly.”[11] Interoperability is crucial not only to ensuring that information displays the same across different computing platforms, but also for metadata harvesting initiatives. This is significant for our collections because commercial search engines have started using metadata harvesting to gain additional access to our resources. For instance, Google implemented the OAI-PMH protocol for acquiring additional data sets, while Yahoo! has acquired content from OAIster.[12]

3. Examples

Metadata is not a singular item that can be quickly added to digital content during its creation. There is a metadata standard and community for every conceivable manifestation of content. What all of these different metadata standards have in common is the concept of the *schema.* At its heart, metadata is *structured information* in which the schema provides the organization by defining the various elements. The descriptive content of an item is defined with a content standard, and the semantic elements that encapsulate the content are defined in either an encoded Document Type Definition (DTD) or Extensible Schema Definition (XSD).[13] The metadata specialists construct these different elements in order to create linked-data. Consider the following snippet of a Dublin Core metadata record encoded in XML:

[9] Berners-Lee, Tim. “Linked Data-Design Issues.” http://www.w3.org/DesignIssues/LinkedData.html. (Accessed: 2010-04-22).

[10] W3C SWEO Community Project. “Linking Open Data.” http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData. (Accessed: 2010-04-22).

[11] NISO, p. 1.

[12] Wikipedia. “Open Archives Initiative Protocol for Metadata Harvesting” http://en.wikipedia.org/wiki/Open_Archives_Initiative_Protocol_for_Metadata_Harvesting. (Accessed: 2010-04-22).

[13] NISO, p. 2-4.

```
<dublinCore>

        <dc:title>Leaf from a fifteenth century French Book of hours
        printed for Simon Vostre.</dc:title>

 </dublinCore>
```

In the above example, the content is: *Leaf from a fifteenth century French Book of hours printed for Simon Vostre*, which was formulated by applying the rules put forth in the Descriptive Cataloging of Ancient, Medieval, Renaissance, and Early Modern Manuscripts (AMREMM.) The semantic element for the above content is the **<dc:title>** component, which is defined by the Dublin Core Metadata Element Set (DCMES) as: “A name given to the resource.”[14]

The DCMES provides the semantics for 15 core elements--the schema “that should be used to describe distributed information resources on the Internet for discovery purposes.”[15] Other metadata standards are far more granular, providing for a level of description that would tax even the most earnest metadata specialist. Complicating matters further is the fact that not all metadata communities use the same core set of definitions to describe their standard, and XML uses many of the same terms for ontologically similar, but fundamentally distinct concepts. Nonetheless, the above general principles regarding metadata as a concept seem to apply fairly well in the library setting.

Before closing this section on basic definitions, it is important to mention that metadata itself can be grouped according to the function it performs in the life-cycle of a digital object. NISO suggests that metadata generally falls under:

- Descriptive metadata: The metadata that describes the resource and facilitates retrieval.
- Structural metadata: The metadata that indicates how compound digital objects are put together.
- Administrative metadata: The metadata concerned with rights, preservation and technical documentation.[16]

All three forms assist in the management of online information and contribute to creating data that can exchange its information with both humans and machines. The term "metadata," therefore, encapsulates a series of concepts beyond those concerned just with description. The degree to which digital objects are used, understood and persist in the online environment depends on the decisions made from the outset of a project to support the different types of metadata. Going forward, this is essential for any project

[14] DCMI. “DCMI Glossary-T” http://dublincore.org/documents/usageguide/glossary.shtml#T (Accessed: 2010-04-22).

[15] Zeng, Marcia Lei and Jian Qin. *Metadata.* New York: Neal-Schuman Publishers, Inc., 2008, p. 6.

[16] NISO, p. 1.

that expects to have an "online" life because: “metadata is key to ensuring that resources will survive and continue to be accessible into the future.”[17]

The next section provides a brief introduction to the intellectual foundations for the organization of information and the rise of metadata communities in the mid-90s.

4. Intellectual Foundations

When discussing metadata, it is important to keep in mind that a core principle is to manage information. This alone should make metadata important to library services. After all, to say that we now live with an overabundance of information is an understatement. At Tufts alone, our collective research has produced an embarrassment of riches, and it is an increasingly herculean task to sift through it all just to figure out who published what, and when. Still, is this overabundance of information fundamentally different than any time in the past? Harvard professor Robert Darnton in the *New York Review of Books* suggests that information has always been difficult to manage, even before the advent of the computer.[18] That the organization of information has always posed challenges is axiomatic to the library profession, which has produced a number of standards and methods for dealing with the large amounts of data it collects. For the past 30 years libraries and cultural institutions have relied almost exclusively on *content standards* such as the Anglo-American Cataloging Rules Revised, Second Edition (AACR2) for describing their collections, and *encoding standards* such as Machine Readable Cataloging (MARC) for encoding their data. The underlying foundations for these standards were a series of intellectual first-principles articulated by Charles Cutter in his Rules for a Dictionary Catalog, which he published in 1876.[19] According to Cutter, information seeking behavior generally starts with a query based on some form of a title, an author, or a subject. As a result, we would do well to remember Elaine Svenonius’ position in The Intellectual Foundation of Information Organization that “the principles, objectives, and techniques that have been developed to organize information within the field of library and information science constitute a body of knowledge with wide application, not the least of which is the organization of information in digital form.”[20] Still, one can’t help but feel that something *has* changed within the digital information environment. A question remains: do the old objectives to locate, identify, select, obtain and even navigate the bibliographic universe still apply in a world now dominated by Web 2.0 technologies that increasingly favor collaborative online learning?

[17] Ibid.

[18] Darnton, Robert. “The Library in the New Age.” The New York Review of Books. 55 (10) 2008, pp. x-x.

[19] Cutter, Charles. *Rules for a Dictionary Catalog, Fourth ed.* Washington. UNT Digital Library. digital.library.unt.edu/ark:/67531/metadc1048/. (Accessed: 2010-04-22).

[20] Svenonius, Elaine. *The Intellectual Foundation of Information Organization*. Cambridge, Mass: MIT Press, 2000, p. xiv.

Unfortunately, much of library literature is silent on the issue of organizing and creating a uniform structure of best practices for collaborative online learning.[21] As a result, information on best practices currently resides within that increasingly authoritative source of information, Wikipedia. A quick glance at the "Wikipedia Manual of Style" reveals that many of the organizing principles first articulated and used by librarians in their content standards is still necessary to corral online information. Consider the following entry under the main heading "capitalization," which instructs users how to contribute information about individuals with titles:

> When used as titles (that is, followed by a name), items such as president, king and emperor start with a capital letter: President Clinton, not president Clinton. The formal name of an office is treated as a proper noun: Hirohito was Emperor of Japan and Louis XVI was King of France (where Emperor of Japan and King of France, respectively, are titles). Royal styles are capitalized: Her Majesty and His Highness; exceptions may apply for particular offices. When used generically, such items are in lower case: De Gaulle was a French president and Louis XVI was a French king. Similarly, three prime ministers attended the conference, but, we know that the British Prime Minister is Gordon Brown. For the use of titles and honorifics in biographical articles, see Honorific prefixes.[22]

This entry is every bit as detailed as anything in the AACR2, along with its commensurate rule interpretation. In fact, the entire "Manual of Style" reads as if it is an updated version of the AACR2 and includes a descriptive rule for every conceivable situation. That a robust online community dedicated to encoding information for easy retrieval essentially reinvented the wheel should indicate that our information organization needs remain relatively unchanged.

Nonetheless, it is the position of this paper that *some* things have changed, and while AACR2 and MARC have performed well in upholding Cutter's Objectives, the fact remains that in the mid-90s metadata development emerged as a concept concomitantly with the popularization of the Internet as a vehicle for the delivery of information. Coupled with the rise of Web 2.0 principles in the last 5 years, it seems appropriate to think about how we want to systematize the creation and delivery of our information to our users. This is especially true as we begin the task of figuring out how user discovery needs, the management of digital rights, and the preservation of digital objects will play into our overall strategic plans as we begin talking about next generation discovery platforms. In order to better understand where we want to go, it is

[21] Haupt, Jon. "From Zero to Wiki: Proposing and Implementing a Library Wiki." Journal of Web Librarianship. Vol. 1 (1) 2007, p. 79.

[22] Wikipedia. "Manual of style." http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style#Titles (Accessed: 2010-04-22)

helpful to consider the history of the metadata communities and the reason for so many different standards.

5. Brief History

Marcia Leng points out that during the mid-90s people managing distributed information repositories started attempting to make sense of what the delivery of information via "the web" actually meant.[23] This was especially true for institutions that managed large amounts of digital information.[24] Internet based information called for completely new "mechanisms of description, authentication, and management, which prompted the development of new guidelines and architectures by different communities."[25] The scientific community, for instance, developed the *Content Standards for Digital Geospatial Metadata* in 1992, while the *Guidelines for Electronic Text Encoding and Interchange* became the de rigueur data standard for the humanities. Meanwhile, the library community began experimenting with metadata for digital objects, eventually resulting in an OCLC sponsored workshop in 1995, and the subsequent emergence of the Dublin Core metadata standard.[26] Since the mid-1990s an alphabet soup of acronyms now adhere to the different metadata standards that dominate the cultural heritage community. These include not only the metadata schemas, but also the XML technologies used to create and encode the resources.[27]

The overwhelming number of standards reflects the simple fact that different metadata communities are attempting to address their specific information needs and audiences. As indicated above, the scientific community created most of the early metadata standards; more recently, visual resource associations and archival communities are creating robust standards to fit their unique needs. In fact, throughout the 90s and early 2000s, metadata standards and their registries grew at a tremendous rate and, oftentimes, without input from related communities. The sheer quantity of metadata standards creates many challenges, not least of which is how to integrate related schemas in order to "describe the same resource for multiple purposes and to serve a number of user groups."[28] Luckily metadata, especially when it is encoded for the web in XML, is extensible, making it possible to create additional metadata schemas to help integration. For instance, the Resource Description Framework [RDF] was developed by the W3C to provide a model for integrating metadata schemas into the description of web related resources.[29] According to NISO:

> In RDF a namespace is defined by a URL pointing to a Web resource that describes the metadata schema that is used in the description. Multiple

[23] Zeng, pp. 6- 8.

[24] Ibid.

[25] Ibid.

[26] Ibid.

[27] Ibid.

[28] NISO, p. 1.

[29] NISO, pp. 11-12.

> namespaces can be defined, allowing elements from different schemas to be combined in a single resource description. Multiple descriptions, created at different times for different purposes, can also be linked to each other. RDF is generally expressed in XML.[30]

The metadata communities that developed in the 90s created a series of standards and methods for combining them that now needs to be integrated into library workflows through the support of metadata services. Needless to say, metadata services require specialists to make sense of, and implement, appropriate standards for different online projects.[31] These individuals not only need to be aware of the different standards for organizing and describing content, but also the technical standards for encoding, such as XML, and its host of attendant technologies like eXtensible Stylesheet Transformations (XSLT), X-Link, and X-Query along with HTML and CSS.

6. Current Developments

Metadata creation and management in a library setting is a complex area of expertise. Marcia Lei Zeng points out that the "coordination of many components is necessary when generating records to ensure the product's quality and interoperability."[32] Although different organizations support different types of metadata services, it is the position of this paper that Tisch Library needs to develop a metadata service for digital content that will support a set of infrastructure components in order to ensure that the metadata produced here is of high-quality, exists at the network level, is linked and is interoperable. This also lays the groundwork for next generation discovery tools.

It is important to begin thinking about metadata in this manner at Tisch, because as Muriel Foulonneau and Jean Riley point out: "The role of metadata within cultural heritage institutions has increased over time, in terms of the functions it provides and the user interactions it enables, as institutional missions have expanded along with the types of material collected. Material that these institutions curate, archive, display and organize is increasingly digital."[33] In fact, at ALA and LYRASIS conferences, and especially within the halls of Simmons Graduate School of Library and Information Sciences, metadata is given increasing prominence because library collections and resources are increasingly online. Jennifer Bowen, Director of Metadata Management at the University of Rochester, and lead developer of the eXtensible Catalog Project, goes so far as to suggest that discussion about next generation discovery platforms are meaningless without:

- an understanding of metadata itself and a commitment to deriving as much value from it as possible;

[30] Ibid.

[31] Foulonneau, Muriel and Jenn Riley. *Metadata for Digital Resources: Implementation, Systems Design and Interoperability.* Oxford, England: Chandos Publishing, 2008, p. 4-6.

[32] Zeng., p. 211.

[33] Foulonneau, p. 4.

- a vision for the capabilities of future technology;
- an understanding of the needs of current (and where possible, future) library users; and
- a commitment to ensuring that lessons learned in this area inform the development of both future library systems and future metadata standards.[34]

The eXtensible Catalog (XC) project at the University of Rochester is an exciting development, as they are laying the ground work for a next generation discovery platform that will use open source applications to "provide libraries with an alternative way to reveal their collections to users."[35] At its heart, is a team which is committed to utilizing the various metadata standards and re-purposing them to meet user expectations. This is not, therefore, a simple quick-fix solution to create a false impression of federated searching and Web 2.0 functionality, but rather a concerted effort to move library data into the network as linked-data. For Bowen, "To present library resources via the Web in a manner users now expect, library metadata must function in ways that have never been required of it before. Making library metadata function effectively within the broader Web environment will require that libraries take advantage of the combined knowledge of cataloging/metadata and system development who share a common vision for serving library users."[36] It cannot be stressed enough that the XC project is one of the most exciting and important developments in the metadata community, and that Tisch would do well to both watch its development, and prepare for next generation metadata services by understanding and using non-MARC metadata now.

7. The Miscellany Collection at Tisch Library

Increasingly, libraries are in the business of producing and managing all types of metadata. Until now, Tisch Library has focused almost exclusively on the creation and maintenance of MARC21 metadata. However, in November 2009, a small collection of medieval and renaissance miscellany was discovered in the Special Collections Department. Traditionally, these types of records would have been cataloged for display in a library OPAC, with a minimal level of description. Needless to say, this is problematic as scholars are starting their research with Google, instead of library catalogs, and collections encoded in MARC21 are largely invisible to web based queries. By encoding the material in XML instead of MARC21, and using the Dublin Core metadata standard, the Miscellany Collection is discoverable to scholars online and also allows visual verification of cataloging information about the items themselves. Scholars may participate in on-line discussions about the project via a custom built Twitter feed, a common Web 2.0 application, and assist in formulating descriptive elements in order to increase access to the collection itself and better meet their research needs. Currently the collection represents 3 records. It will eventually grow to

[34] Bowan, Jennifer. “Metadata to Support Next Generation Library Resource Discovery: Lessons from the eXtensible Catalog, Phase 1” Information Technology and Libraries. 27 (no 2) June 2008, p. 1.
[35] Ibid.
[36] Ibid., p. 16.

33 records and 58 associated images. The records themselves will be encoded as linked-data, utilizing the RDF framework to increase both interoperability and discoverability. This is also important because RDF is seen as the basic framework for the *Semantic Web*, or Web 3.0. The *Semantic Web* is envisioned by Sir Tim Berners-Lee and the World Wide Web Consortium (W3C) as: "a vision of information that is understandable by computers, so computers can perform more of the tedious work involved in finding, combining, and acting upon information on the web."[37] As the web shifts to this new paradigm for organizing information, metadata will become even more important. Consequently, libraries must start utilizing the appropriate encoding, content and technical metadata standards for all online data presentation, or else their efforts will become completely invisible in the Web 3.0 world. Rather than repeating the mistake of the British National Archives, Tisch Library is attempting through *The Miscellany Collection* pilot project to design a proof of concept for creating linked metadata for library collections using the appropriate standards for the current web. It is important for Tisch Library to support these projects if we are truly serious about our vision statement: "Tisch Library connects people to information at Tufts and beyond": a vision statement that practically presupposes networked data, and its metadata underpinnings.

[37] Wikipedia. “Semantic Web.” http://en.wikipedia.org/wiki/Semantic_web. (Accessed: 2010-04-22).