Merging Metadata: A Sociotechnical Study of Crosswalking and Interoperability

Michael Khoo, Catherine Hall
The iSchool at Drexel University
3141 Chestnut Street
Philadelphia, PA 19104, USA
+1 215 895 1230
{khoo, ceh48}@drexel.edu

ABSTRACT
Digital library interoperability relies on the use of a common metadata format. However, implementing a common metadata format among multiple digital libraries is not always a straightforward exercise. This paper reviews some of the metadata issues that arose during the merger of two digital libraries, the Internet Public Library and the Librarian’s Internet Index. As part of the merger, each library’s metadata was crosswalked to Dublin Core. This required considerable work. A sociotechnical analysis suggests that the metadata for each library had been shaped in complex ways over time by local factors, and that this complexity negatively impacted the efficiency of the crosswalk. Some implications of this finding for digital library interoperability are discussed.

Categories and Subject Descriptors
H.3.7 Digital Libraries – collection, standards, user issues.

General Terms
Management, Standardization

Keywords
crosswalk, Dublin Core, interoperability, metadata, operations, organizational knowledge, organizations, sociotechnical

1. INTRODUCTION
Interoperability allows digital libraries to share resources, catalogs and search queries amongst each other, and to reach wider audiences. There are a number of technical prerequisites for successful interoperability, including the use of standardized metadata. With some digital libraries, this standardization is built directly into those libraries’ infrastructure. In other cases, such as digital libraries that have developed independently with no history of cooperation, but which wish to share resources, catalogs and users, interoperability has to be developed from the ground up.

This paper examines digital library interoperability from a sociotechnical perspective. The term ‘sociotechnical’ is used here in an inclusive sense to cover a range of theoretical approaches that claim, broadly, that social and technological phenomena are mutually constitutive: that is, that technological processes shape social processes, while social processes simultaneously shape technical processes. Examples of theoretical approaches include the science technology and society (STS) approach of Bijker [5]; the social informatics approach of Kling [19]; and studies informed by the structurational theory of Giddens [11]. Sociotechnical approaches to the study of digital libraries assume that they are composed of multiple social, technological, and other phenomena, linked together in mutually constitutive ways which can generate emergent, complex and sometimes unpredictable outcomes [30].

2. Metadata
Metadata is used to describe and index digital library resources [16]. Good quality metadata supports users to search and retrieve relevant resources, while poor quality metadata hides resources and negatively affects user satisfaction. [2, 3, 10]. Interoperable metadata allows the users of one library to search and retrieve resources from more than one library [1, 7, 27]. The adoption of a common metadata format (such as Dublin Core) opens the way to partnerships between libraries that either have adopted this format, or who are willing to crosswalk - that is, to map their existing metadata - to that standard format. Again, good quality metadata supports interoperability, while poor quality metadata inhibits interoperability.

A range of metadata quality issues has been identified. These include issues related to the quality of the metadata in a particular field, such as completeness (the number of fields utilized) and accuracy (the correctness of the data entered), as well as structural metadata issues such as differences in elements between the two different schemas [e.g. 2, 3, 23, 25, 31, 33]. In the context of interoperability, Bruce and Hillman [7] describe seven dimensions of metadata quality - completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. Stvilia et al. [28] provide four categories for metadata sharing and federated collection development - semantic consistency, structural consistency, completeness, and ambiguity.

2.1 Sociotechnical Metadata Issues
Metadata, like the digital libraries in which it is used, is sociotechnical in nature. When analyzing metadata quality, sociotechnical factors therefore have to be taken into consideration. Several broad overlapping groups of sociotechnical metadata factors have been described, including logistical issues, tool issues, knowledge and communication issues, and organizational issues.

First, the number of resources requiring cataloging is growing faster than the number of catalogers available to do this work. This has resulted in a growing metadata ‘bottleneck’ between metadata needs and metadata capacity [21]. Consequently,
metadata may be generated by catalogers who are not trained in metadata theory and practice, and this in turn can affect metadata quality. Second, while a variety of tools exists to support the creation of metadata for digital libraries, their usability can be affected by issues such as interface design, inadequate support documentation, and tool learning curves, again affecting metadata quality [8, 9, 13, 16, 18]. Third, there can be knowledge and expertise issues that affect the ease or difficulty with which metadata knowledge is shared within a project, affecting metadata operations [17]. Fourth, metadata is often crafted in local contexts to address particular institutional and professional needs, [4, 6, 22, 29] and subsequent crosswalking out of these local formats and into a common standard can be problematic.

3. The IPL-LII Merger

This paper examines the impact of these sociotechnical metadata issues in the context of the merger of the Internet Public Library (IPL: http://www.ipl.org/) with the Librarian’s Internet Index (LII: http://www.lii.org/; this URL is no longer live). Part of the merger involved the crosswalking of both libraries’ metadata to Dublin Core. As will be described, the crosswalk required more time and effort than had been anticipated. We argue that sociotechnical factors played an important part in these delays. (The authors were members of a fluctuating voluntary roster of approximately 12 faculty, graduate students, and technical staff, which addressed the development of the back end, front end, and services for ipl2 as necessary. The following account is reconstructed from personal observations and notes, internal project email and documentation, and interviews with various project team members).

The IPL was created in 1995. By 2010, it provided authoritative collections, information assistance, and information instruction for the public, and was available throughout the United States as a training tool for library and information science (LIS) programs. It had subject-categorized collections of more than 40,000 online resources, generated by students, volunteers, and staff members [14]. In 2007 it was decided to develop the next version of the IPL, ‘ipl2.’ As part of this development process, it was decided to crosswalk existing IPL metadata to Dublin Core (DC). Quality metadata was thought to be one foundation upon which future Web 2.0 user-centered services (such as user accounts, and tagging and annotation) could be developed.

The IPL had also developed a partnership with the California-based Librarian’s Internet Index (LII). The LII subsequently lost its funding, and it was decided to merge the two libraries as part of the IPL re-launch. The prior adoption of DC by the IPL supported this move, as it provided a framework within which LII metadata could be crosswalked to DC and then ingested into ipl2.

Many unexpected metadata issues were encountered during the merger, including the following. (In this discussion, the original IPL metadata is referred to as ‘IPL metadata’; the original LII metadata is referred to as ‘LII metadata’; and the new Dublin Core metadata is referred to as ‘ipl2 metadata.’)

3.1.1 Unique or specialized metadata fields

A number of elements in IPL and LII metadata were highly granular. For instance, IPL metadata included a set of fields related to the resource title, including Main Title (a good match for dc:title), but also Former Title, Sort Title, Acronym, Alternate Title, and Alternate Spelling. Most of these were rarely used, and in general there was no organizational memory as to why these sub-elements existed. The discovery of these and other sub-elements in the IPL and LII metadata precipitated an extended series of discussions. On one hand, it was felt that the extra metadata was not really relevant to ipl2; on the other hand, it was thought that the metadata had been created by a human being and must therefore say something useful about the resource. In the end, much of this metadata was placed in custom administrative fields, archived ‘out of sight’ of ipl2 users.

3.1.2 Lack of item-level metadata

During the merger it was realized that many IPL collections had collection-level records, but no item-level records for items in those collections. IPL users were finding these pages as a result of using third party search engines (Google, Yahoo!, etc.) and not the IPL search engine. This also meant that there would be no metadata for these resources to be crosswalked to Dublin Core. Collection-level records were present (however incompletely) primarily because it was necessary to create such a record in order to create the collection itself, and to make it appear on the IPL website. There was however no obligation to create item-level records for the resources that were in these collections. The reasons why item-level metadata was not added are now unknown, and the organizational memory associated with this decision had disappeared.

3.1.3 Collections stored in different databases

It was assumed prior to the crosswalk that all the IPL metadata was located in one place (a MySQL database). It was discovered however that at least two of the special collections in the IPL were housed in separate Filemaker Pro (FMP) databases. These collections could not be included in the original crosswalk. One explanation was that when the IPL originally began operation, the original catalogs were stored in FMP format, but that as the IPL grew and became successful, some of the catalog - but not all of it - was migrated to MySQL. The collections left in FMP format are now being manually crosswalked to Dublin Core. This slow process that is still ongoing, and until it is completed, these collections will not be available in IPL2.

3.1.4 Lack of controlled subject vocabularies

Both the IPL and LII metadata included a ‘subject’ element. However neither library consistently used controlled vocabularies for this element. Subject headings in the IPL were ‘usually taken from the IPL’s existing subject headings or from the Sears List of Subject Headings’ (cataloging documentation; emphasis added), but in practice a controlled vocabulary was not strictly applied. In the case of the LII, Library of Congress Subject Headings were used with some collections, but not consistently over time. The LII used to employ a cataloger who could apply LCSH terms, but this position was discontinued when funding was reduced, and subsequent catalogers did not adopt LCSH. Practically speaking, there was no controlled vocabulary for ‘subject’ either in the IPL or the LII - and thus also in ipl2.

3.1.5 Lack of hierarchy in browse structures

Both the IPL and LII allowed users to browse resources by subject. In the IPL, for example, users could click on high-level subject categories in a left hand navigation column to pull up a dynamic menu that could be used to drill down through subject sub-categories to particular collections or resources. In both the IPL and LII users could also browse by clicking on links on a series of web pages that again led them deeper into the collections in each library. The path the user took was recorded in a
‘breadcrumbs’ style navigational trail dynamically generated at the top of each browsing page.

One problem that emerged well into the crosswalk process was the realization that the browse structures were not hierarchical. Some of the LII browsing metadata contained ‘loops,’ where ‘parent’ and ‘children’ subject categories were mutually defined as parents and children of each other. Children could also be linked to more than one parent category. Especially in the case of the LII, it was possible to assume that one was drilling down into the collection, when in fact one was moving to another unrelated branch in the browse structure, or even back up to a higher category. Similar problems (although to a lesser degree) were found in the IPL browsing structure. Once again, there was no explanation for the existence of these discrepancies, beyond the assumption that this was partly an outcome of piecemeal library development and a lack of organizational memory.

3.1.6 Incommensurate subject categories

The browse structures of each library did not map onto each other. For instance, Egyptian pyramids were classified under ‘History’ in IPL, and under multiple subjects (‘Geography,’ ‘Archaeology,’ ‘Architecture,’ etc.) in the LII. Before the merger, it had been assumed that it would be possible to take resources from the LII browse hierarchy, and place them in at least roughly equivalent positions in the IPL hierarchy. However, this was not possible, as at times this would have meant placing large numbers of resources from a high-level LII subject category, into a particular and fine-grained IPL sub-category that was intended only to have a few resources in it. Here, the mis-matches seem to have arisen as a consequence of the incremental growth of both the IPL and the LII resulting in libraries beginning with a relatively small number of subject headings, then being expanded to include other subject headings; and as the original subject headings became populated, catalogers in both libraries were encouraged to add one or more new subject headings, but not in a controlled fashion.

3.1.7 Complex metadata workflow

The IPL metadata administration tool was developed over time and from the ground up. It was a functional but specialized assemblage of web pages, online forms, and front end and back end hacks. It was necessary to develop a new metadata tool to administer the new Dublin Core schema. Just as the original IPL metadata turned out to be more complex than expected, so did the IPL metadata tool, with its complicated structure and permissions regarding who could read, edit, approve, etc., metadata records. It was necessary to develop a new cataloging tool from scratch, but opaque understanding of how the old tool functioned hindered this work.

4. SOCIOTECHNICAL DIMENSIONS OF METADATA CROSSWALKING

A non-sociotechnical analysis of the IPL-LII merger might see the delays described in this paper as arising from a series of technical issues requiring individual attention which, when addressed, would culminate in the merger of the IPL and LII. However, these issues were also embedded within a wider sociotechnical context, including issues arising from unrecognized and often undocumented legacy systems and decisions, whose rationale had faded into the background. These sociotechnical issues were hard to identify, either because they had become tacit and ‘taken for granted’ (c.f. [24, 29, 32]) and therefore do not seem unusual or noteworthy, or because they had been forgotten completely.

Addressing these issues resulted in significant work for the merger team. It was often difficult to understand how and why a particular aspect of the IPL metadata took the form it did, which in turn made it difficult to understand how the issue could be ameliorated. An interesting aspect of this situation was that while the IPL team encountered these issues on a regular basis, many of them only became apparent after the crosswalk process was underway; whereas one might have assumed that it would have been advantageous to have been able to identify these idiosyncrasies at the start of the project. One explanation for this is that both the IPL and LII metadata had been developed in sociotechnically specific circumstance that had subsequently been forgotten.

This preliminary finding has practical implications for digital library interoperability. Digital libraries have much to gain from federating; however this is often difficult to achieve, especially for digital libraries with limited resources [27]. One important outcome of this ongoing work is therefore the research question: What forms of organizational knowledge tools would usefully and practically support the management of digital library metadata and interoperability work, especially for digital libraries with limited resources?

5. CONCLUSION

This paper has offered a sociotechnical exploration and explanation of legacy metadata issues and their effect on digital library mergers and interoperability. It has demonstrated how the local, specific, and often pragmatically ad hoc nature of metadata work generates metadata formats and structures that ‘bind’ metadata to those points of origin, with the consequence that much work is then needed to ‘unbind’ them again, for instance during a crosswalk. This characteristic of metadata is relevant for any organization engaged in crosswalking or adopting standardized metadata. Such an activity can turn out to be more resource-intensive than anticipated; and there is scope for the development of practical organizational strategies and tools to support digital libraries, especially those with limited resources, to engage in metadata crosswalking and interoperability activities.

6. ACKNOWLEDGMENTS

We would like to thank the Drexel iSchool faculty, staff and doctoral students involved in this work. Part of the work was supported by a grant from the OCLC-ALISE Library and Information Science Research Grant Program for 2009.

7. REFERENCES


Wenger, E. Communities of Practice. New York: Cambridge University Press.