Addressing the Metadata Bottleneck by Developing and Evaluating an Online Tool to Support Non-specialists to Evaluate Dublin Core Metadata Records

Michael Khoo
The iSchool at Drexel University
michael.khoo@ischool.drexel.edu

Final Report
2009 OCLC/ALISE Library and Information Science Research Grant Project

July 12, 2011

Suggested citation:
Khoo, Michael. 2011. "Addressing the Metadata Bottleneck by Developing and Evaluating an Online Tool to Support Non-specialists to Evaluate Dublin Core Metadata Records." 2009 OCLC/ALISE research grant report published electronically by OCLC Research. Available online at:
http://www.oclc.org/research/grants/reports/2011/khoo2011.pdf

1. Why the research is innovative in the context of current related research

The funded research has addressed and continues to address fundamental issues in the design of metadata tools for non-experts. Metadata work has traditionally been the preserve of specialists; however, the explosive growth in the World Wide Web has resulted in a growing gap between the number of online resources and limited numbers of professionals who can create metadata for them. There is therefore a 'metadata bottleneck' that restricts metadata production and has negative consequences for the quality and utility of online repositories. This bottleneck is a central research problem for metadata, digital libraries, and the expansion of information infrastructure. It can be addressed in different ways. More metadata specialists could be trained and hired, although this is unlikely. Automatic metadata generation is another option, but progress here varies; it is relatively easy to extract metadata from homogeneous resources, such as articles written in the same format, but it is harder to deal with more heterogeneous resources, such as web pages. New forms of metadata (such as tagging and folksonomies) are a third possibility, although there are concerns here regarding the quality and accuracy of the tags generated. A fourth strategy involves the use of non-specialists (such as resource creators) in structured metadata work. These include technical metadata creators, such as data in-putters, who generally have had basic training, but have not participated in a formal educational program; content creators, who create metadata for intellectual content they have authored, who are amateurs, professionals, and the average citizen; and public metadata creators, community or subject enthusiasts who create metadata for a resource authored by other individuals (Greenberg, 2010). The original research proposal called for the development of a metadata tool to support such non-professional users in the creation and editing of Dublin Core metadata for online collections for the Internet Public Library (IPL).

In 2008, work began on the next version of the IPL. The aim was to develop a new look for the site and to begin providing Web 2.0 services such as user accounts and personal collections. Part of the work involved crosswalking IPL metadata from its existing format to Dublin Core, and moving the metadata from the existing MySQL database to a new Fedora database. Dublin Core was chosen to make sharing IPL metadata with other libraries easier, and because it was perceived to offer simplicity and flexibility in terms of elements and definitions. As part of this work, a new metadata administration tool for Dublin Core was required. Existing metadata tools have had a mixed track record, partly related to the success with which they convey complex metadata concepts to non-specialists. Poor interface usability, tools that do not represent metadata concepts in a clear user-centered fashion, and other problems, can hinder metadata work.

The proposal therefore advocated a user-centered design approach to improve metadata tool quality and usability for the IPL (Crystal and Greenberg, 2005; Greenberg et al., 2003; Wilson, 2007). The research involved in this approach would contribute to a better understanding of the design of such tools, support non-professionals to create Dublin Core metadata, and thus help to address the metadata gap.

2. Scope and methodology employed for the research, including completed phases of the research with dates

As part of this user-centered design approach, and informed by theories of organizational knowledge, the research gathered data on the tacit and explicit knowledge of the various groups involved in metadata work in the IPL. The aim was to find out how metadata work was carried out in a real life organizational context, and to build the metadata tool in accordance with the needs of organizational members. The

research focused on gathering project documents, analyzing organizational communication, and interviewing project members. In the original proposal, the data was to be analyzed partly through Centering Resonance Analysis, a latent semantic analysis tool. Progress in this particular area was limited, as the development of this tool had been stopped and it was hard to get hold of the developers.

At the same time, a wider series of issues emerged with the IPL metadata crosswalk, including: lack of item-level metadata for resources in many of the collections; rarely utilized and arcane metadata fields and sub-fields; confusing subject browsing hierarchies; database legacy issues; and tool interface and workflow issues. These issues were very complex, and they had to be resolved before the crosswalk could be implemented, and before work on the metadata tool itself could proceed.

The research focus of the project therefore switched to an investigation of existing difficulties with the metadata crosswalk, and how these could be addressed, and therefore inform the design of the metadata tool. It was thought that this would be a relatively quick exercise, but the research turned out to be time-consuming.

3. The analyses and discuss the findings of the research

IPL project members were interviewed about the metadata difficulties in the project. The interviews were transcribed, and the transcripts were coded. The analysis of the interviews revealed that the work of making and sharing metadata in the IPL was very complicated. It was tightly-coupled work, that is, "work that strongly depends on the talents of collections of workers and is nonroutine, even ambiguous. Components of the work are highly interdependent. The work typically requires frequent, complex communication among the group members, with short feedback loops and multiple streams of information" (Olson & Olson, 2000). Factors such as the presence of different forms of organizational tacit knowledge in the IPL, all affected the ways in which this work was carried out, and the problems that were encountered. Significant organizational communication had to be carried out with and between the various groups in the IPL project (such as faculty, students, and IT staff) in order to make sure that the issues were addressed. As the resulting organizational processes were to be embodied and described in the new metadata tool, they therefore ultimately contributed to the design of this tool.

4. Outcomes of the research

The 'Erasmus' metadata tool for Dublin Core cataloging[1]

While it would have been relatively easy to build a metadata tool for the IPL, to build a tool that was acceptable to all the members of the project, and which was also easy to use, was considerably harder. However, after a number of problems and delays - including the problems with the Dublin Core crosswalk, programmers leaving, upgrades in web browsers breaking parts of the site, and other issues - the emphasis on organizational communication in the IPL project, partly based on the analyses developed in (3) above, led to the development of a prototype of a novel but useful tool for metadata work.

The components of the tool itself are modular and extensible, and can be edited and reconfigured through interacting with a back-end database. This design makes it possible easily to develop build and test a cataloging tool for any 'flavor' of Dublin Core with relatively little effort. Each element of any

[1] *And although Ptolemy's library was confined within the narrow walls of his dynastic palace, Aldus toils so that his library shall be contained by no limits other than those of the world. (*Desiderius Erasmus, *Festina Lente* ('*Make Haste Slowly*'), 1525.)

The IPL's Erasmus metadata tool is named after the Dutch humanist and philosopher Desiderius Erasmus, who in 1525 wrote of the potential of a new information technology, the printing press, to break down physical limitations to the spread of knowledge and to create a 'library without walls.' Almost five centuries later this dream is yet to be fully realized. At the same time, there continues to be a rapid development of digital information technologies that claim to be able to support such a library.

schema, and its description, can be entered into the back end, and the tool will then generate a front-end interface for that particular schema on-the-fly every time that particular schema is used. The schema configuration can then be saved for re-use, and can also re-edited at any time. We currently have a working prototype of the tool, which will be subject to initial user testing over the summer and fall of 2011. We hope to have a working beta version of the tool ready for demonstration at the 2012 ALISE conference in Dallas, Texas (and a version of this report will be submitted to this conference as a full abstract). Once the tool is stable, we will distribute a full description of its architecture. We are committed to the continuing development of the tool in the context of the IPL.

A sociotechnical model of metadata work in organizations

In addition to the tool itself, the work has also resulted in the generation of a useful sociotechnical model of metadata work in organizations (Khoo & Hall, *under review*). The model emphasizes the strong dependencies that metadata work has on apparently non-metadata organizational activities, such as database design and web site design. It emphasizes the importance of bringing all project members to 'the same page' in order to discuss and address these differences. This model was used successfully to plan the development of the new IPL metadata tool just described.

Relevant papers, presentations and publications

Galloway, M., M. Khoo, X. Lin, X., & J-r. Park. (2009). Crosswalking IPL metadata to Dublin Core. iConference 2009, University of North Carolina at Chapel Hill, NC, February 8-11, 2009.

Khoo, M. (2010). Addressing the 'Metadata Bottleneck' by Developing and Evaluating an Online Tool to Support Non-specialists to Evaluate Dublin Core. ALISE 2010 Conference, January 12-15, 2010, Boston MA.

Khoo, M., & Hall, C. (2010). Merging Metadata: A Sociotechnical Study of Crosswalking and Interoperability. 10th ACM/IEEE Joint Conference on Digital Libraries, Brisbane, Australia, June 21-25, 2010, pp. 361-364.

Khoo, M., & Hall, C. (2011). Surprise, Surprise: Organizational Dimensions of Metadata Work in an Educational Digital Library. *Under review.*

Khoo, M., X. Lin, & J-r. Park. (2009). *A User-Friendly Metadata Quality Control Tool for the Internet Public Library*. 9th ACM/IEEE Joint Conference on Digital Libraries (JCDL), Austin, TX, June 15-19, 2009, pp. 407-408.

Khoo, M., Park, J-r., and X. Lin. (2009). The User-Centered Design Of A Non-Specialist Metadata Tool And Interface For The Internet Public Library. 72nd Annual Meeting of the American Society for Information Science and Technology, Vancouver, BC, November 6-11, 2009.

5. Suggestions for future research based on the results of the project

The tool is still in prototype stage, but we hope to have a stable alpha version available over the summer, which will be made available for IPL interns to start using in initial usability testing. Based on these usability results, further iterations of the tool design may be made over the coming year. Recommendations for metadata tool design will also be generated. In addition, opportunities will be sought for reapplying the sociotechnical model of metadata work generated during the research, in different organizational settings.

6. References

Greenberg, J. (2010). Metadata and digital information. In M. J. Bates & M. N. Maack (Eds.), *Encyclopedia of Library and Information Sciences, Third Edition* (pp. 3610-3623). New, York, NY: Routledge.

Olson, G. M., & Olson, J. S. (2000). Distance matters. *Human-Computer Interaction 15,* 139-178.