

Metadata: A Promising Solution

A few years ago, metadata was spoken of only in relation to “library-like” collections. Today, however, most organizations are facing the challenges presented by growing institutional digital repositories and are looking to determine whether metadata might provide a solution to these challenges. As the 2003 Dublin Core Conference stated the premise of its meeting: “Metadata is fundamental to persons, organizations, machines, and an array of enterprises that are increasingly turning to the Web and electronic communication for disseminating and accessing information.”¹ Basically, metadata is fundamental to everyone!

Although an industry appears to be growing up around metadata, no empirical evidence or ROI analysis has made a definitive case that metadata is the solution to the problem that metadata is intended to address: that of access to information—usually in the form of reports, proposals, technical documents, meeting minutes, etc.—within an organization’s digital repository. Anecdotal evidence suggests that using metadata in an intranet environment can reduce the time that employees spend looking for, finding, and verifying files, but it remains to be seen in a large-scale implementation and evaluation whether these savings and other benefits are sufficient to offset the current cost of producing metadata for items in a repository. Little information exists about the full costs. In one example, Mike Doane, a consultant with SBI and Company, reported that his company typically charges from \$195,000 to \$275,000 to initially set up a metadata solution for a corporation.² After that, the

corporation faces the ongoing costs of metadata assignment to all new documents—and this is where the bulk of the costs lie.

But must the ongoing process of metadata assignment be expensive? Not necessarily. Basically, there are three ways for metadata to be assigned to items as they are added to a repository: (1) manual assignment by trained, dedicated personnel; (2) self-assignment by the document producers; or (3) automatic metadata generation. Although the first two of these methods carry significant costs in terms of personnel time, the third offers a promising new approach for generating timely, cost-efficient, effective metadata.

For instance, the Center for Natural Language Processing at Syracuse University has developed and tested technology—MetaExtract—that automatically generates and assigns metadata to electronic documents.³ In one application, metadata was automatically assigned to documents such as lesson plans from the educational field. To understand the size of this task, note that in the standard metadata schema for educational resources, there are twenty-four elements that the system needs to assign: Audience;

Keywords; Cataloger; Contributor; Language; Duration; Coverage; Publisher; Educational Level; Creator; Identifier; Quality; Date; Document Type; Standard; Description; Rights Management; Relation; Essential Resources; Source; Subjects; Format; Title; Pedagogy. In the MetaExtract system, natural language-processing technology utilizes the multiple levels of human language processing for the information-extraction task of recognizing appropriate values for all of the twenty-four elements that are present in the document and for producing a metadata record based on these extractions as a rich representation of the content of the document.

Experimental evaluation of the quality of the metadata produced by MetaExtract and of the performance of the system in



Photo Illustration by David Bishop, © 2005

user tasks of finding relevant resources in the educational domain has demonstrated that when compared with manually assigned metadata, the automatically assigned metadata performed well, providing more consistent, broader coverage of the elements from the schema and equal quality both in the minds of the users and in retrieval of relevant resources. Furthermore, potential system users who served as subjects indicated that the metadata could perform multiple functions: retrieving relevant resources, providing a quick and rich summary of the resource itself, and enabling easy browsing across sets of retrieved results.

Given the demonstrated ability for automatic metadata generation in the domain of educational resources, the general issue to be addressed is whether this capability can be extended to—and evaluated as performing equally well for—digital repositories in other domains. Experience in the highly diverse fields of public health⁴ and aerospace engineering⁵ suggests that in fact automatic metadata generation is as viable for these fields as for the educational domain and is therefore likely to be equally extensible to other domains and genres.

The development of a successful automatic metadata-assignment solution for a new domain requires either the adoption of an appropriate preexisting metadata schema or the elicitation from users of the repository of the necessary information content elements for a useful metadata schema. Specialist communities reveal a nonproductive tendency to see their data as unique and to resist the notion that a general strategy could inform their work. Groups that are considering the adoption of a metadata solution for their digital repository are strongly encouraged to build on the experiences and successes of the active metadata community. Even though there are usually several unique elements that need to be added to represent documents of a particular domain or genre, standardized metadata schema such as the Dublin Core Metadata Initiative (<http://www.dublincore.org/>) or the Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMH (<http://www.openarchives.org/>), are highly likely to provide a basic, preliminary core set of metadata elements. Building on such

well-accepted and widely used metadata schema will pave the way for future interoperability across archive repositories.

Metadata offers a promising solution to institutional needs for providing improved access to important information contained in a wide range of documents. However, manual metadata assignment, like cataloging, is a labor-intensive and costly function, requiring special knowledge and training. With the ever-increasing quantity of materials requiring such attention, the need for efficient metadata assignment is all the more keenly felt. Interestingly, the resulting tension between “efficient” metadata assignment and “quality” metadata assignment⁶ may in fact be artificial, given the preliminary evidence that high-quality metadata can be automatically assigned quite efficiently, even if a human reviewer is included in the loop.

Notes

1. 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice—Metadata Research & Applications, September 28–October 2, 2003, Seattle, Washington, <<http://dc2003.ischool.washington.edu/>>.
2. Mike Doane, “Metadata, Search, and Meaningful ROI,” Global Corporate Circle DCMI 2003 Workshop, Seattle, Washington, September 28, 2003, <<http://dublincore.org/groups/corporate/Seattle/Circles-Workshop-Papers/DC2003-Doane.ppt#328,2,Slide 2>>.
3. Ozgur Yilmazel, Christina M. Finneran, and Elizabeth D. Liddy, “MetaExtract: An NLP System to Automatically Assign Metadata,” Center for Natural Language Processing, School of Information Studies, Syracuse University, <http://www.cnlp.org/presentations/slides/JCDL_2004_Final.ppt>.
4. A. M. Turner, E. D. Liddy, J. Bradley, and J. A. Wheatley, “Modeling Public Health Interventions for Improved Access to the Grey Literature,” *Journal of the Medical Library Association* (forthcoming).
5. E. D. Liddy, “Extraction of Elusive Information from Text,” International Association of Science and Technology for Development (IASTD) International Conference on Knowledge Sharing and Collaborative Engineering, St. Thomas, U.S. Virgin Islands, November 22–24, 2004.
6. Thomas R. Bruce and Diane I. Hillmann, “The Continuum of Metadata Quality: Defining, Expressing, Exploiting,” in Diane I. Hillmann and Elaine L. Westbrooks, eds., *Metadata in Practice* (Chicago: American Library Association, 2004).

Elizabeth Liddy is a Trustee Professor in the School of Information Studies at Syracuse University and Director of its Center for Natural Language Processing, where she leads a team of researchers focused on developing human-like language-understanding software technologies.



EDUCAUSE

Transforming Education Through Information Technologies

EDUCAUSE, a consolidation in 1998 of Educom and CAUSE, is a nonprofit consortium of colleges, universities, and other organizations, dedicated to the transformation of higher education through the application of information technologies. Through direct services and cooperative efforts, EDUCAUSE assists its members and provides leadership for addressing critical issues about the role of information technology in higher education.

EDUCAUSE Board of Directors

Perry O. Hanson, Chair

CIO and Associate Provost for Educational Technology
Brandeis University

Kathleen Christoph, Vice Chair

Director, DoIT Academic Technology Solutions
University of Wisconsin–Madison

Robyn R. Render, Secretary

Vice President for Information Resources and CIO
University of North Carolina, Office of the President

John E. Bucher, Treasurer

Director of Information Technology
Oberlin College

John C. Hitt

President
University of Central Florida

Rebecca L. King

Director for Information Systems & Services
Baylor University

Jeffrey W. Noyes

CIO and Associate Vice President for Information Technology
University of Texas at San Antonio

Margaret F. Plympton

Vice President for Finance & Administration
Lehigh University

David L. Smallen

Vice President, Information Technology
Hamilton College

George O. Strawn

CIO
National Science Foundation

Ellen J. Waite-Franzen

Vice President, Information Services
Brown University

David Ward

President
American Council on Education

Ex Officio Member

Brian L. Hawkins
President
EDUCAUSE