User Generated Metadata: Creating Personalized Web Experiences

Michael Fegan *(mfegan@msu.edu)*
Matrix, Michigan State University
Bill Hart-Davidson *(hartdav2@msu.edu)*
WIDE Center, Michigan State University
Joy Palmer *(palmerjo@msu.edu)*
Matrix, Michigan State University
Dean Rehberger *(rehberge@msu.edu)*
Matrix, Michigan State University

Abstract

This session will focus on the importance of measuring how communities of users interact with digital objects. By drawing on user-performance data and metadata generated for secondary repositories, we explore ways to enhance use and access of documents and digital libraries..

Speaker 1 proposes an approach to representing the structure of a document based on the way readers or users interact with it in the context of a deliberative task. This approach contrasts with other ways to model the structure of documents including approaches which map authorial intention and those which rely upon a well-known information model or genre. This presentation will highlight the benefits of understanding the structure of documents based on the rhetorical reading/using strategies of those who interact with them. These structures can be rendered as 'paths' through a given set of information resources, offering insight into the way that objects and relationships that make up a document can mediate, or complicate, activity. Speaker 1 will conclude by showing some examples which make use of user-performance data to create task-appropriate views of complex (multiscale) documents.

Speakers 2 and 3 will examine the role of secondary repositories can play in enhancing access and interaction for students and scholars in the humanities. The most entrenched *a priori* models for information structuring and delivery online are derived from library and archival cataloguing practices. In line with digital library best practices, digitized sources are typically cataloged to describe their bibliographic information, along with technical, administrative, and rights metadata. While these practices are essential for preserving the digital object and making it available to users, unfortunately they do so in a language and guise often difficult to understand within the context of use. In addition, materials in digital libraries do not literally 'speak' for themselves and impart wisdom; they require interpretation and analysis within a context of use. Access and use of digital objects can no longer be thought of in terms of stand alone files or individual digital objects, but rather must directly impact the ways in which users reuse, repurpose, combine and build complex digital objects. This assumption relies on a more complex meaning for the term *access* that will be detailed and explained in this paper.

Following the examples in the first paper, speaker 4 will demonstrate an application that can be used to collect user generated metadata. Following the concepts developed in the second paper, speaker 4 will develop the argument in practice that one way we can enhance access to online digital objects is to facilitate the creation of secondary repositories. These repositories will provide discipline/community specific metadata and applications and will allow users to find, use, manipulate and analyze digital objects more easily. To this end, Speaker 4 has developed *Media Matrix* 1.0 — an online, server-side suite of tools that allows users to locate specific media and streaming media files found in digital repositories and segment, annotate and organize this media online. This application provides users with an environment both to work with and personalize digital media, and also to share and discuss their findings with a community of users. This paper will explore if the creation of secondary repositories of usage statistics and user-generated materials/metadata (to supplement both traditional cataloging records and discipline-specific online indexes) can help scholars and students in the humanities gain better access to online materials.

Modeling Documents Based on User Performance: An Alternative to Author Intention and a priori Information Model Approaches

Bill Hart-Davidson, Ph.D.

This paper proposes an approach to representing the structure of a document based on the way readers or users interact with it in the context of a deliberative task. This approach contrasts with other ways to model the structure of documents including approaches which map authorial intention and those which rely upon a well-known information model or genre. This presentation will highlight the benefits of understanding the structure of documents based on the rhetorical reading/using strategies of those who interact with them. These structures can be rendered as 'paths' through a given set of information resources, offering insight into the way that objects and relationships that make up a document can mediate, or complicate, activity. The paper will conclude by showing some examples which make use of user-performance data to create task-appropriate views of complex (multiscale) documents.

There is a great deal of interest today in the idea of building models of texts. One reason is that, with the growth of the Web as a way to reach a wide and diverse audience, publishers of information of many types are now interested in building information structures that support multiple-audience adaptation. Another reason is to maximize the value of content by delivering information that is tailored to a particular task.

For example, imagine the day-to-day work of a claims processing agent for a large insurance company. The agent is responsible for making decisions based on information in documents of various types - claim forms, telephone call records, police and adjustor reports, medical records, even photographs - all stored in electronic policyholder files. There may be discernable patterns in these types of workflows which can be documented and used as the basis for information models that structure information at or below the level of an individual document. The model could allow information contained in all the documents associated with the policyholder file to transform to suit the decision-making needs of the specific users who interact with it.

Creating a system like the one described above would require analyzing and modeling fundamental patterns of document use, defining a basic modeling language for document-mediated interaction that can capture recurrent patterns of user performance. As Moser & Moore (1996) point out, most semantic modeling approaches construct formal text structures based on either author intention or, alternatively, on an information structure presumed to be instantiated in the text. Neither of these approaches is entirely appropriate for creating effective displays of information for potential users. Creating such displays requires a model of the text-mediated interaction between writers and readers which can then be used to define display conditions for a range of information "views" that a given document might support.

Performance-based text structure models differ from other types in that they are not primarily representations of a stable 'core' semantic structure that is assumed to be either domain independent (e.g, Mann and Thompson, 1987), or genre specific, as suggested by the work of Bazerman(1988) and others. Nor are these structures maps of author intentionality and/or struggles in creating intentional relationships similar to analyses by Van Wijk and Sanders (1999). Rather, the models emphasize structures that constitute the resources authors and the specific users or readers of a document share in order to come to some kind of agreement about an issue or question that all parties have a stake in. What gains status as a 'unit' or 'object' in user-performance based models depends upon the deliberative activity that the document is meant to support. Relationships among objects are similarly defined by how the objects mediate a given decision. In this way, we can expect the model to account for both the regularities in text structures which correspond with similar texts doing similar mediational work, as well as quite specific and arbitrary text structures associated with any given deliberative activity as it unfolds in a social context. This modeling approach comes very close to a process described by Phelps (1985) who articulated an approach to structural analysis drawing on and responding to work by Faigley & Witte (1981) and Van de Kopple (1985) in composition studies, as well as Halliday & Hassan (1976) and Van Dijk (1976) in linguistics. The process, broadly, understands texts as objects with histories, requiring us to study them' in process' if we are to understand how they shape the experiences of a reader.

Bibliography

Bazerman, C. *Shaping written knowledge.* Madison: University of Wisxonsin Press, 1988.

Faigley, L., and L.S. Witte. "Coherence, cohesion, and writing quality." *College Composition and Communication* 32 (1981): 189-204.

Halliday, M.A.K., and J.R. Hassan. *Cohesion in English.* London: Longman, 1976.

Mann, W.C., and S.A. Thompson. "Rhetorical structure theory: Toward a functional theory of text organization.." *Text* 8.3 (1988): 243-281.

Moser, M., and J.D. Moore. "Toward a synthesis of two accounts of discourse structure." *Computational Linguistics* 22.3 (1996): 409-419.

Phelps, L.W. "Dialectics of coherence: Toward an integrative theory." *College English* 47.1 (1985): 12-31.

Van Dijk, T. *Text and context: Explorations in semantics and pragmatics of discourse.* London: Longman, 1976.

Van Wijk, C., and T. Sanders. "Identify writing strategies through text analysis." *Written Communication* 16.1 (1999): 51-75.

Vande Kopple, W. "Given and new information and some aspects of the structures, semantics, and pragmatics of written texts." *Studying writing: Linguistic perspectives.* Ed. C.R. Cooper and S. Greenbaum. Beverly Hills, CA: Sage, 1986. 72-111.

Enhancing Access to Online Digital Objects through Reciprocity between Primary and Secondary Repositories

Dean Rehberger and Joy Palmer

This paper will examine the role of secondary repositories can play in enhancing access and interaction for students and scholars in the humanities. While access to online resources has steadily improved in the last decade, online archives and digital libraries still remain difficult to use, particularly for students and novice users (Arms). In some cases, a good deal of resources have been put into massive digitization initiatives that have opened rich archives of sources to a wide range of users. Yet, the traditional cataloging and dissemination practices of libraries and archives make it difficult for these users to locate and use effectively these sources, especially within scholarly and educational contexts of the humanities. Many digital libraries around the country, large and small, have made admirable efforts toward creating user portals and galleries to enhance the usability of their holdings, but these results are often expensive and labor intensive, often speaking only directly to a small segment of users.

To address these problems, we begin with the assumption that access and preservation are mutually dependent concepts. Preservation and access can no longer be thought of in terms of stand alone files or individual digital objects, but rather must directly impact the ways in which users reuse, repurpose, combine and build complex digital objects. This assumption relies on a more complex meaning for the term *access*. Many scholars in the field have called for a definition of access that goes beyond search interfaces to the ability of users to retrieve information "in some form in which it can be read, viewed, or otherwise employed constructively" ((Borgman 57)). Access thus implies four related conditions that go beyond the ability to link to a network:

1. equity the ability of 'every citizen' and not simply technical specialists to use the resources;
2. usability the ability of users to easily locate, retrieve, use, and navigate resources;
3. context the conveyance of meaning from stored information to users, so that it makes sense to them;
4. interactivity the capacity for users to be both consumers and producers of information.

The keys to enhancing access for specific user groups, contexts, and disciplines are to build secondary repositories with resources and tools that allow users to enhance and augment materials (Shabajee), share their work with a community of users (Waller), and easily manipulate the media with simple and intuitive tools (or at least build interfaces that match existing, well-known applications). Users will also need portal spaces that escape the genre of links indexes and become flexible work environments that allow users to become interactive producers (Miller).

Herbert Van de Sompel has proposed a successful system (OpenURL/SFX framework for context sensitive reference linking) for disaggregating reference linking services from e-publishing. In his framework, the service of providing links between references and across e-publisher's digital repositories is separated from the services provided by the e-publishers. In so doing, the service provides "seamless interconnectivity between ever-increasing collections of heterogeneous resources" , freeing primary repositories from the difficult and expensive task of ensuring links to references while giving users greater access to resources and increasing the value of the digital object (Van de Sompel). Similarly, we propose the concept of secondary repositories that would be responsible for handling secondary metadata, extended materials and resources, interactive tools and application services. This information is cataloged, stored, and maintained in a repository outside of the primary repository that holds the digital object. The comments and observations generated by users in this context are usually highly specialized because such metadata is created from discipline-specific, scholarly perspectives (as an historian, social scientist, teacher, student, enthusiast, etc.) and for a specific purpose (research, publishing, teaching, etc.). Even though the information generated by a secondary repository directly relates to digital objects in primary repositories, secondary repositories remain distinctly separate from the traditional repository. The information gathered in secondary repositories would rarely be used in the primary cataloging and maintenance of the object, and primary repositories would continue to be responsible for preservation, management, and long-term access but would be freed from creating time-consuming and expensive materials, resources, services, and extended metadata for particular user groups.

In line with digital library best practices, digitized sources are typically cataloged to describe their bibliographic information, along with technical, administrative, and rights metadata. While these practices are essential for preserving the digital object and making it available to users, unfortunately they do so in a language and guise often difficult to understand within the context of use (Lynch 2003). Even though the author's name, the title of the work, and keywords are essential for describing and locating a digital object, this kind of information is not always the most utilized information for ascertaining the relevance of a digital object. For instance, K-I2 teachers often do not have specific authors or titles in mind when searching for materials for their classes. Teachers more frequently search in terms of grade level, the state and national standards that form the basis of their teaching, or broad overarching topics derived from the required content and benchmark standards (e.g., core democratic values or textbook topics) that tend to display too many search returns to make the information of value.

While cursory studies have indicated these access issues, still very little is known about archival use or how these users express their information needs (Duff, Duff & Johnnson). For digital libraries to begin to fulfill their potential, much research is needed to better understand the processes by which primary repositories are accessed and how information needs are expressed. For example, research needs to address the ways in which teachers integrate content into their pedagogy so that bridges can be built from digital repositories to the educational process, bridges that greatly facilitate the ability of teachers and students to access specific information within the pedagogical process. Recent research strongly suggests that students need conceptual knowledge of information spaces that allow them to create mental models to do strategic and successful searches. As with any primary source, the materials in digital libraries do not literally 'speak' for themselves and impart wisdom; they require interpretation and analysis lysis (Bowker & Star; Duff; Duff & Johnson). Allowing communities of users to enhance metadata and actively use, reuse, repurpose, combine and build complex digital objects can help users to contextualize the information they find, draw from deeper resources within the digital library, and find more meaningful relationships between digital objects and their needs. Thinking in terms of a distributed model (similar to the open source software community) that allows users both easier access to materials and a greater range of search criteria and also provides opportunity for active engagement in the generation of metadata and complex digital objects, promises to help us rethink our most basic assumptions about user access and long-term preservation.

Collections can also benefit by defining communities of users. For example, with the recent release of secret White House tapes (<http://millercenter.virginia.edu/>), the sheer number of tapes and hours make it impossible for adequate cataloging of content as well as the difficulty of determining the context and people involved (or even what is said given the poor quality of many tapes). Those historians and scholars (a more regulated and highly defined set of experts) allowed access to the collections could supply information about content and context as well as set terms for debates over more questionable areas of interpretation (e.g., when sound quality makes passages inaudible). While metadata gathered in these ways would need to be qualified (maintained in a secondary repository) because of lack of quality control, the processes could make large quantities of data that is key to many disciplines in the humanities more available and usable.

Bibliography

Arms, William. *Digital Libraries.* Cambridge, MA: MIT Press, 2001.

Borgman, C. *From Gutenbeg to the Global Information Infrastructure: Access to Information in the Networked World.* Cambridge, MA: MIT Press, 2000.

Bowker, Geoffrey C., and Susan Leigh Star. *From Gutenbeg to the Global Information Infrastructure: Access to Information in the Networked World.* Cambridge, MA: MIT Press, 1999.

Cole, Charles. "Name Collection by Ph.D History Students: Inducing Expertise." *Journal of the American Society for Information Science* 51.5 (2000): 444-455.

Cooperstock, J.R. "Classroom of the Future: Enhancing Education through Augmented Reality." *Proceedings of the HCI International, Conference on Human-Computer Interaction, New Orleans.* 2001. Accessed 2005-04-15. <http://www.cim.mcgill.ca/~jer/pub/hcii01.pdf>

Duff, Wendy. "Evaluating Metadata at a Metalevel." *Archival Science* 1 (2001): 285-294.

Duff, Wendy, and Catherine A. Johnson. "A Virtual Expression of Need." *American Archivist* 64 (2001): 43-60.

Hedstrom, Margaret. "Research Challenges in Digital Archiving and Preservation." NSF Post Digital Libraries Futures Workshop. 15-17 June 2003.

Kornbluh, Mark Lawrence, Dean Rehberger, and Michael Fegan. "Media MATRIX: Creating Secondary Repositories. in Research and Technology for Digital Libraries." *Proceedings of the 8th European Conference, ECDL2004.* Berlin: Springer, 2004. 329-340.

Lynch, Clifford. "Reflections toward the Development of a ‘Post D-L Research Agenda." NSF Post Digital Libraries Futures Workshop. 15-17 June 2003.

Lynch, Clifford. "Interoperability: the standards challenge for the 90s." *Wilson Library Bulletin* March (1995): 38-42.

Lynch, Clifford. "Colliding with the Real World: Heresies and Unexplored Questions about Audience, Economics, and Control of Digital Libraries." *Digital Library Use: Social Practice in Design and Evaluation.* Ed. Ann Bishop, Barbara Butterfield and Nancy Van House. Cambridge, MA: MIT Press, 2001. 191-218.

Marshall, Catherine. "Annotation: from Paper Books to the Digital Library." *Proceedings of the ACM Digital Libraries ‘97 Conference, Philadelphia, PA.* July 23-26, 1997. Accessed 2005-04015. <http://www.csdl.tamu.edu/~marshall/dl97.pdf>

Miller, Paul. "The Concept of the Portal." *Ariadne* 30 (20 December 2001). Accessed 2005-04-15. <http://www.ariadne.ac.uk/issue30/portal/intro.html>

Page, K.R., D. Cruickshank, and D.D. Roure. "Its About Time: Link Streams as Continuous Metadata." *Proceedings of the Twelfth ACM Conference on Hypertext and Hypermedia (Hypertext '01).* 2001. 93-102.

Rydberg-Cox, Jeffrey. "Cultural Heritage Language Technologies: Building an Infrastructure for Collaborative Digital Libraries in the Humanities." *Ariadne* 34 (14 January 2003). Accessed 2005-04-15. <http://www.ariadne.ac.uk/issue34/rydberg-cox/intro.html>

Shabajee, Paul. "Primary Multimedia Objects and 'Educational Metadata': a Fundamental Dilemma for Developers of Multimedia Archives." *D-Lib Magazine* (March 2000). Accessed 2005-04-15. <http://www.dlib.org/dlib/june02/shabajee/06shabajee.html>

Van de Sompel, Herbert, and O. Oren Beit-Arie. "Open linking in the scholarly Information Environment Using the Open URL Framework." *D-Lib Magazine* (March 2001). Accessed 2005-04-15. <http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html>

Waller, Richard. "Functionality of Digital Annotation: Imitating and Supporting Real-World Annotation." *Ariadne* 35 (30 April 2003). Accessed 2005-04-15. <http://www.ariadne.ac.uk/issue35/waller/>

Developing MediaMatrix: A Secondary Repository Tool

Michael Fegan

Speaker Two will argue that one way we can enhance access to online digital objects (particularly in the humanities) is to facilitate the creation of secondary repositories. These repositories will provide discipline/community specific metadata and applications and will allow users to find, use, manipulate and analyze digital objects more easily.

Even though access by specialist scholars and educators to digital objects has grown at an exponential rate, tangible factors have prevented them from fully taking advantage of these resources in the classroom, where they could provide the conceptual and contextual knowledge of primary objects for their students. When educators do find the materials they need, using objects from various primary repositories to put together presentations and resources for their students and research can be challenging. Beyond merely creating lists of links to primary and secondary resources, assembling galleries of images, segmenting and annotating long audio and video files require far more technical expertise and time than can realistically be expected in the educational context. In addition, even though scholars have a long history of researching archives and are comfortable sifting through records, locating items, and making annotations, comparisons, summaries, and quotations, these processes do not yet translate into online tools. Contemporary bibliographic tools have expanded to allow these users to catalogue and keep notes about media, but they do not allow users to mark specific passages and moments in multimedia, segment it, and return to specific places at a later time. Multimedia and digital repository collections thus remain underutilized in education and research because the tools to manipulate the various formats often 'frustrate would be users' and take too much cognitive effort and time to learn.

To this end, Speaker Two has developed *Media Matrix* 1.0 — an online, server-side suite of tools that allows users to locate specific media and streaming media files found in digital repositories and segment, annotate and organize this media online. The application has been developed as part of the *Spoken Word Project* funded by Digital Libraries Initiative II: Digital Libraries in the Classroom Program, National Science Foundation in conjunction with UK's Joint Information Systems Committee.

This application is an online tool that allows users to easily find, segment, annotate and organize text, image, and streaming media found in traditional online repositories. *MediaMatrix* works within a web browser, using the browser's bookmark feature, a familiar tool for most users. When users find a digital object at a digital library or repository, they simply click the *MediaMatrix* bookmark and it searches through the page, finds the appropriate digital media, and loads it into an editor. Once this object is loaded, portions of the media can be isolated for closer and more detailed work — portions of an audio or video clip may be edited into a time-segment, images may be cropped then enlarged to highlight specific details. *MediaMatrix* provides tools so that these media can be placed in juxtaposition, for instance, two related images, a segment of audio alongside related images and audio, and so forth. This can be particularly effective for students and researchers who need to fit images into a presentation or would like to demonstrate specific nuances and details about portions of images or artwork. Most importantly, textual annotations can be easily added to the media, and all this information is then submitted and stored on a personal portal page.

A portal page might be created by a scholar-educator who wishes to provide specific and contextualized resources for classroom use, and/or by a student creating a multimedia-rich essay for a class assignment. While these users have the immediate sense that they are working directly with primary objects, it is important to emphasize that primary repository objects are not actually being downloaded and manipulated. *MediaMatrix* does not store the digital object, rather, it stores a pointer to the digital object (URI) along with time or dimension offsets the user specified for the particular object and the user's annotation for that particular object. This use of URI pointing as opposed to downloading is especially significant because it removes the possibility that items may be edited and critiqued in contexts divorced from their original repositories, which hold the primary and crucial metadata for such objects.

As long as primary repositories maintain persistent URIs for their holdings the pointer to the original digital object will always remain within the secondary repository, which acts as a portal to both the primary collection and contextualizing and interpretive information generated by individuals on items in those collections. This information can be stored in a relational database along with valuable information about the individual, who supplies a profile regarding their scholarly/educational background, and provides information of the specific purposes for this work and the user-group (a class, for example) accessing the materials. *Media Matrix* is a PHP based server side application that stores information in a *mySQL* database and exports that information into XML for display. The development of the tool and programming environment have been designed to keep it library and archive independent so that it can work with almost any site on the internet. It can also work easily with any of the standard courseware packages. The tool is also search independent because it relies on traditional internet search tools and a site's discovery tools to find an object. Once objects are found, *Media Matrix* is deployed by the user. Because *Media Matrix* does not actually copy the digital object from the site (it only stores a pointer to the object in the form of a URI and whatever time offsets are created by the user), it avoids some of the copyright and fair use pitfalls that often keep users from working with digital objects (although there are issues of deep linking to be addressed). The secondary repository can thus be searched and utilized in any number of ways.

Historians, for example, can browse the portals of other historians working specifically in their research areas or K-12 teachers can browse grade appropriate sections defined by specific grade levels and subjects to see what digital objects other teachers are using or, more important, for time challenged teachers, they can find specific presentations created around standard topics and curriculum frameworks. Users can also perform keyword searches over the annotations created by all users or specific groups of users. A teacher, for instance, can choose to search through only the information in eleventh-grade Civics groups in hopes of finding information that speaks directly to his/her needs. Because users have gathered content from across the Internet and from a variety of digital repositories, searching *Media Matrix* is equivalent to searching multiple repositories at once. Once users find an object from a particular digital library, they can jump to that repository to find what other objects are available.

Going beyond demonstration, this paper will also dive the latest findings and evaluations based on initial user testing in several classrooms as Tufts University and Michigan State University.