

Digital Letters

Winter 2003-04

Issue Number Four

Letter from the Editor

Hello, and welcome to the fourth issue of *Digital Letters*. This issue focuses on something very near and dear to my heart—metadata. Surprised? In trying to decide what to write about in this issue it was only natural that I chose to write about something that's a significant part of my work here at UCSD. As a metadata analyst for the Union Catalog for Art Images (UCAI) project I review metadata for consistency and how well it conforms to a standard. The quality of our metadata directly affects whether our end users will be able to successfully search, retrieve, and utilize the resources they need. Hopefully this issue will help illuminate its complexity and the issues involved in dealing with metadata.

In addition, this issue will give you a flavor of some conferences that UCSD librarians attended this past fall. In their own words, they'll discuss sessions and issues that left lasting impressions on them. You'll notice ALA midwinter is absent from this issue but look for coverage of this important conference in the Spring issue. If you would like to write up a brief summary of any digital library sessions (i.e. any innovative ways that technology is being applied in libraries today) from ALA midwinter or any other conferences you've attended recently please let me know at trose@ucsd.edu. I welcome all submissions. Submissions for the Spring issue will be accepted until March 30, 2004.

Enjoy! ~ Trish Rose

METADATA: MORE THAN JUST "DATA ABOUT DATA"

Introduction

Metadata plays a key role in today's digital libraries. The term "metadata" is widely used both inside and outside the library community but its most common definition – "data about data" – does little to express its full potential and value. More than just a surrogate for a resource, metadata will help ensure digital resources can be managed, preserved, more widely distributed and accessible in the future. Priscilla Caplan, Assistant Director for

(Continued on page 2)

Reports from the Field

Fall of 2003 was a particularly busy one for UCSD librarians and staff who attended conferences in addition to their regular duties, including the American Society for Information Science and Technology (ASIST) conference in Long Beach, CA during October 19th - 22nd and the Digital Library Federation's fall forum in Albuquerque, NM, November 17th– 19th.

ASIST

The theme of this year's conference was Humanizing Information Technology which inspired a lot of sessions dealing with social informatics – how people use and interact with information and the user in terms of their information-seeking behavior. At least 3 UCSD librarians and one staff attended in addition to a former UCSD librarian, Amy Wallace, who won the ASIST Leadership Award for 2003. Overall, the UCSD attendees felt it a very worthwhile conference. Below are highlights from two UCSD librarians about what they found most interesting.



When you design a kitchen, do you bring in an interior designer or do it yourself? Will the designer understand your needs better than you? Can the designer help you build a better kitchen? If the answers to these questions are obvious to you, consider the design of a web site, a database, or an online catalog. Do we, as information professionals, have a better sense of how an information system should be designed or should we ask the users?

User-centered (or not) design was one of the hottest topics in this year's ASIST conference as evidenced by several panel discussions devoted to it. As one of the speakers put it, although user-centered design has been of relevance to system designers for decades, a renewed interest on the topic has reached a "critical mass" due to the ease of web page creation. While some speakers thought that user studies must be conducted to design better systems, others argued that good design would attract its own users and that the user-centered approach has led to more problems because there are no "normal users." Others believe designs will fail if one relies solely on users for input. While the arguments did not end there, someone from the audi-

(Continued on page 3)

METADATA *(continued from page 1)*

Digital Library Services at the Florida Center for Library Automation states that its two essential components are a) that it describes a resource and b) that it should be composed according to a documented metadata schema.¹ In order to get a better grasp on the metadata concept it may be easier to think of it according to the three primary functions that it serves: descriptive/analytical; administrative/preservation; and structural.

Descriptive metadata, as its name implies, is used to describe a resource and put it into context with other resources. Examples of descriptive metadata include title, author/creator, description, and subject. In this sense, metadata is serving the same purpose as cataloging has for bibliographic resources, namely access, but it has a more expanded role than cataloging because electronic documents present new challenges over paper resources relating to management, long term access, and structural representation.

According to Paul Conway, Director of Information Technology Services at Duke University Libraries,

“the digital world transforms traditional preservation concepts from protecting the physical integrity of the object to specifying the creation and maintenance of the object whose intellectual integrity is its primary characteristic.”²

Digital resources, if they are to be available in the future, require extensive data about their technical creation (e.g. software needed for use, date of creation, image resolution); rights (i.e. in what manner a resource can be used); and provenance (e.g. date and time of updates, information needed for media migration).

Complex digital resources that are comprised of several simple digital objects, such as images and text from an illustrated book or a multimedia presentation, also need to have data recorded about their structural organization in order to display and help end users navigate those resources.

Metadata schemas

As we've learned from the bibliographic world, in order for metadata to be useful, it must conform to a recognized standard, such as a schema. A metadata schema (sometimes referred to as a data dictionary) is essentially a set of metadata elements or fields and rules for their use. While schemas vary widely depending on the communities they are created for they can include: data elements and semantics; content rules; and syntax.

At a minimum, a metadata schema should provide a list of elements and their definitions or semantics. The schema should also indicate whether each element is required, optional, or conditionally required and whether the element is repeatable (useful if, for example, a resource has more than one creator). If necessary, the schema may also specify hierarchical relationships between the elements (e.g. parent/child). Content rules specify how data should be selected and recorded within the elements (e.g. use the fullest form of name that you know beginning with last name first). The name could then be recorded either as separate elements separating first and last names or within a single element. The syntax specifies how elements should be encoded in machine-readable form so that data can be more easily exchanged between programs and systems. An example would be eXtensible Metadata Language (XML).³

As digital resources have increased so have the proliferation of schemas. Different communities or domains have different information needs. Dublin Core is a schema designed to be the “lowest common denominator” that can be used across communities and for any type of data. While there is value in having such a universal schema in order to bring together heterogeneous data, there can be significant data loss by “dumbing down” all the data to a common set. A few of the many domain schemas familiar to the library community include: VRA Core (for visual resources); TEI (for text documents); and EAD (for archival documents). It is not as important, or perhaps even possible, for all communities to agree to use the same schema because they have different needs. It is far more essential to be able to map across schemas (a.k.a. crosswalking) in order to enable data exchange and distribution.

Challenges in metadata mapping

Whether migrating data from legacy systems or intending to share and distribute resources more widely, metadata will inevitably have to be mapped to different standards over time. How consistently the metadata rules have been applied to the data is the most important factor in determining the difficulty of such a mapping. Beyond that, other mapping issues will be encountered to varying degrees.

Source and target schemas may have different levels of granularity such as when data in one schema may need to be mapped to more than one element in another. For example, an element called Style/Period in one schema may need to be split into separate Style and Period elements in another, requiring manual intervention in order to determine what information is appropriate for each element. Mapping may also involve tradeoffs between data loss and compression. For example, when there is no element in the target schema with an equivalent meaning the data may either be left behind or put in a more general element (e.g. the latitude/longitude of where a picture was taken could be transferred to a general note element).

(Continued on page 4)

Reports from the Field

(Continued from page 1)

ence summarized it nicely: any design, whether it is the design of a kitchen or a web site, is a science. The design of an information system follows a set of principles and requires talent and training to do it right. However, we, as information professionals, do not know it all. If we design an inflexible system, users will by-pass us and go directly to Amazon or Google.

– SuHui Ho, *Digital Services Librarian, S&E Library*

As a government information librarian, I found this year's ASIST conference to be extremely useful even though I was only able to attend for one day. Our own Leigh Star (from the UCSD Communications Department) gave one of the plenary talks about the social construction of information systems. Those familiar with her work, *Sorting things out: classification and its consequences* – a collaboration with Geoffrey Bowker – would recognize her recurring themes of ethnographies in computer science, the sociology of technology, and classification theory.

One of the first sessions of the day, "Transborder data flow: implications for information dissemination and policies between the United States, Canada, and Mexico" featured Nadia Caidi (University of Toronto) and Pierrette Bergeron (University of Montreal). Caidi and Bergeron discussed general information policy in the age of globalization as well as specifics of the Canadian government's policies and their Government Online Initiative (GOL). Especially relevant and interesting was the discussion of information as a commodity versus information as a service and the current trend toward trade and commodification within the context of the World Trade Organization (WTO), General Agreement on Tariffs and Trade (GATT), and the WTO's Trade-Related aspects of Intellectual Property Rights (TRIPS) agreement. It was also educational to hear of a national government (Canada) creating robust delivery of web-based, bi-lingual information and services to its citizens and being conscious about global access standards and services to those with physical and sensory disabilities.

An afternoon session, "Public domain under pressure" by Dr. Paul Uhlir (U.S. National Research Council), Dr. Subbiah Arunachalam (M S Swaminathan Research Foundation), and Tom Moritz (American Museum of Natural History), gave a third world perspective on the growing commercialization of information, and the public domain in a global information economy. While most of what was discussed – the importance of a strong public domain for academic research, the open access movement, open-source software, the Public Library of Science's new open access journal, Wikipedia and Sourceforge – was already well-known to this librarian, it was heartening to hear the discussion of these important issues at such a well-attended session. It is clear to me that the public domain can only survive if these issues are investigated and publicly discussed. Academics and political leaders need to be made aware of its importance.

–James R. Jacobs, *Government Information Librarian, SSSL Library*

DLF

UCSD attendees at this fall's DLF Forum included Esme Cowles and Linda Barnhart, both who reported on the UCAI project, as well as Dawn Talbot and Chris Frymann. Talbot summarizes her thoughts below.



About 160 registrants heard reports on topics ranging from tools and architectures being developed for digital library initiatives to digital resource management, escholarship and user needs. The plenary session delivered by Michael Keller described plans to create a distributed open digital library, known as DODL. It will be a collaborative digital library providing wide electronic access to collections in multiple institutions without assembling those collections in one place, i.e. a virtual library collection. DODL will begin with humanities and social sciences resources that are openly accessible and will incorporate numerous service layers, including an extensive search and discovery service. Next steps include: raising money; recruiting a coordinator; and forming both collections development and technical infrastructure working groups. At the plenary session it was also announced that since membership in DLF has been extended to institutions outside the U.S., expressions of interest have been received from several institutions with extensive experience in digital library development.

Long-term preservation of digital resources was the focus of several sessions. Reports were given on development of a digital format registry, JSTOR's format validation tool, implementation of the LOCKSS program, and costs associated with preservation of digital video. The always entertaining, Herbert Van de Sompel, outlined repository architecture being developed at LANL using MPEG-21 DIDL (Digital Item Declaration Language), OAI-PMH and Open URL, and Clay Shirky spoke about the developing architecture model for LC's NDIIPP soon to be simplified in name at least to NPP.

Sessions devoted to resource management included an update of the DLF E-Resource Management Initiative, which is determining what functionality and metadata are required to enable librarians to manage electronic resources over time, and from Stanford a suite of tools for processing, preserving, and delivering numeric data from social science collections - the Data Extraction Web Interface System (DEWI).

In other sessions, representatives of the University of Chicago described their use of Greenstone digital library software to create a digital collection of musical scores. Cornell discussed their development of a new system for finding digital materials and gave one of the best descriptions of Fedora (Flexible Extensible Digital Object Repository Architecture) through which repositories manage and deliver multiple kinds of digital content. In a non-technical session, librarians from the University of Washington recounted what they learned from a retreat with faculty members who experiment with digital media in scholarship and teaching. Out of the retreat came proposals that the library create a center for digital scholarship to provide support services and that the university create a degree-granting institute to support and study digital scholarship.

Forums for 2004 will be in New Orleans and Baltimore. For DLF's summary of the forum: <http://www.diglib.org/forums/fall2003/Forum-Nov03summ.htm> To view presentations from the forum: <http://www.diglib.org/forums/fall2003/fallforum03.htm#p5>

For info on DLF Fellowships for librarians new to the profession (within 3 years) to attend future forum's see: <http://www.diglib.org/forums/fall2003/fellowship.htm>

METADATA *(Continued from page 2)*

While challenging, mapping data elements can seem like child's play when compared to the stickier task of reconciling heterogeneous data content. Data content usually becomes more of an issue when your records will be searched along with records from another institution. If for example one institution recorded its creator as *Bernini, Giovanni Lorenzo* and the other as *Bernini, Gian Lorenzo* an end user may have a difficult time retrieving all the records for that creator. You essentially have two choices - to reduce all variants to a single preferred form (a.k.a. normalizing) or develop "synonym rings" to connect the variant forms so the user can find the resource regardless of the form they use.

Bringing it all Together

So, now that all of this metadata has been recorded for a digital object you still need a way to bring the metadata and digital content file together into a single unit or package. This type of digital packaging is more commonly referred to as "digital wrapping" and is a method for "binding digital content files and their related metadata together and for specifying the logical relationship of the content files"⁴ The Metadata Encoding and Transmission Standard, better known as METS, acts as just that sort of wrapper. METS is used for encoding descriptive, administrative, and structural metadata and associating it with its corresponding digital content. Since we've run out of space in this issue to go into great detail about METS, look for a more in-depth treatment of this topic in the next issue of *Digital Letters*.

¹Caplan, Priscilla. *Metadata Fundamentals for All Librarians*. Chicago: American Library Association, 2003: 3. (available in the UCSD SSH library - Z666.5 .C37 2003)

²Conway, Paul. *Preservation in the Digital World* Washington, DC: Commission on Preservation and Access, 1996. <http://www.clir.org/pubs/reports/conway2/index.html>

³Note: The terms syntax and schema are used interchangeably in some contexts. Besides the more general definition of a schema that is used in this article, the term schema is also used to refer to a type of XML mark up or encoding. XML encodings were originally created as DTDs – legacy implementations from the SGML world. XML DTDs are slowly being replaced by XML Schemas which have more flexibility. Some XML Schemas include - LC's MARCXML <http://www.loc.gov/standards/marcxml/> and Stanford's XMLMARC <http://laneweb.stanford.edu:2380/wiki/medlane/xmlmarc>. MODS is an XML schema based on MARC <http://www.loc.gov/standards/mods/>. TEI, EAD, and VRA Core do not currently have XML Schemas.

⁴From Online Archive of California definitions page <http://www.cdlib.org/inside/projects/oac/bpgdo/definitions.html>

More about Metadata

For general overviews of metadata:

The Getty's *Introduction to Metadata*

http://www.getty.edu/research/conducting_research/standards/intrometadata/

For further discussion of crosswalks:

Issues in Crosswalking Content Metadata Standards Margaret St. Pierre and

William P. LaPlant, Jr. <http://www.niso.org/press/whitepapers/crswalk.html>

Online maps/crosswalks:

The Getty has created several crosswalks between primary metadata schemas within the library community at http://www.getty.edu/research/conducting_research/standards/intrometadata/3_crosswalks/index.html

UKOLN as also created numerous crosswalks <http://www.ukoln.ac.uk/metadata/interoperability/>

For some of the most up-to-date metadata topics check out these blogs:

<insilico/> <http://libserv12.princeton.edu/insilico/>

Catalogablog <http://catalogablog.blogspot.com/>

For the metapage of all metadata pages see: <http://www.ifla.org/II/metadata.htm>