

Digital Letters

Summer 2005

Issue Number Eight

Letter from the Editor

Welcome to Issue Eight of *Digital Letters* (DL). You may notice DL has a new look—several requests for changes to make DL a bit easier on the eyes resulted in font and layout changes. Thanks to input from Linda Barnhart, James R. Jacobs, Ardys Kozbial, and Rob Melton. I welcome your thoughts on the new look and feel of DL.

This issue features an interview with Brad Westbrook on a metadata standard being applied at UCSD for managing and storing our digital assets. We also have an update from James R. Jacobs on how our UCSD blog, *diglet*, has evolved since James' last article two years ago. In addition, Jenny Reiswig gives us highlights from the survey sent out from the Digital Communications Group in January.

Hope you enjoy!

~ Trish Rose, Image Metadata Librarian, UCAI

METS: A Data Standard for Access and Preservation Now and Into the Future

This month I sat down with Brad Westbrook, Metadata Librarian at UCSD, to talk about METS, the Metadata Encoding & Transmission Standard (<http://www.loc.gov/standards/mets/>). In his position as a Metadata Librarian for the UCSD Libraries, Brad is primarily responsible for developing models and assisting with establishing policy for managing the UCSD Libraries' digital assets (sometimes referred to as digital objects or resources) that are currently being ingested and stored in DSpace - what we now refer to as our Digital Asset Management System (DAMS). Brad has advocated for the use of a metadata standard called METS for that purpose.

TR: Brad it appears METS will play a primary role in managing our digital assets so I was hoping you could give us a better understanding of what METS is all about and how it will serve us in the UCSD Libraries.

BW: I'd be happy to. METS is an XML standard that is a type of digital wrapper. It functions to relate the components of a digital resource. When a

(continued on page 2)

The New and Improved *diglet*

Yes, that's right. *diglet* is new and improved! *diglet* was started by Steve Lawson in April, 2002. For the last 2 years, *diglet* has been a labor of love for Jim Jacobs, Trish Rose and myself. But now, the writer base has been expanded to include the entire Digital Library Program Working Group (DLPWG) including Dawn Talbot, Ken Calkins, Chris Frymann, Gabriela Montoya, Jenny Reiswig, and Brad Westbrook.

What's that you say? What or who is this "diglet" character and what is she doing skulking around our library?

diglet (pronounced "dij-lit") is a blog. According to Wikipedia (<http://en.wikipedia.org/wiki/Blog>), a blog is, "a web application presented as a webpage consisting of periodic posts, normally in reverse chronological order." A blog is a website which is frequently updated with information collected from other websites, news services, listservs etc. as well as comments, thoughts and off-the-top-of-our-heads remarks - in other words, a combination clipping service and sticky notes on Web steroids.

As the online persona of the DLPWG and *Digital Letters*, *diglet* tracks the many issues affecting the digital library landscape including: conferences and workshops; copyright and digital rights management (DRM); metadata and cataloging; open access; digital library technologies; preservation and archiving; papers and articles of interest; and the tables of content of the growing number of journals which focus on digital library issues.

diglet is not only meant for the UCSD library, but is open to the entire Web public. You might not know it from looking, but *diglet* receives on average over 300 hits per day from around the world (some from Google and other search engine spiders, but many from real live people interested in the same things in which we're interested!). So stop by today and find out what's happening with digital libraries! <http://gort.ucsd.edu/mtdocs/diglet>

~James R. Jacobs, Government Information Librarian, SSHL

METS (continued from page 1)

book, for instance, is digitally scanned, its contents are dispersed across many files, one for each page scanned. To assure the integrity of the overall object and to facilitate the use of it, the structural relationship of these files needs to be captured. In addition, a digital wrapper serves to bind metadata and the digital assets together for easier management and storage in a repository or DAMS. A digital asset can be a single file type or a mixture of file types (ex. image, audio, video, or text file). Currently, libraries are using METS to record both presentation and preservation information so that digital assets can be accessed and presented both now and in the future. Some libraries are emphasizing the presentation aspect, using METS to assure the digital assets are displayed in a certain manner. Other libraries are emphasizing the preservation aspect, using METS to capture rich technical and rights metadata for supporting effective management over the long run. The two strategies are not mutually exclusive.

TR: As an international metadata standard who is the responsible party for the development and dissemination of METS?

BW: Primary oversight for METS is provided by the METS Editorial Board which is comprised of librarians and technologists from North America, Australia, and Europe. It has prospered from support from the Digital Library Federation (DLF), and their "public documents" are available on a website provided by the Library of Congress (LoC).

TR: Who are the big players in METS?

BW: Large research libraries who can be categorized into two primary groups: 1) those working on tool development for production and presentation and 2) those working on building METS collections. Those working on tools would include NYU and its work on METS schema and profiles; Harvard and its work on METS production tools and metadata extraction tools; CDL which has developed best practice guidelines for digital objects; and Indiana University, now developing a METS Navigator. METS collection builders include LoC and Harvard, as well as CDL. As far as I know, there is no formal count of METS collections available.

UCSD has played a role in tool development by virtue of its collaboration with CDL on the best practices guidelines for digital objects. Also, UCSD is positioned to soon have one of the larger collections of METS objects. The digital images from AAL will be transformed into 200,000 METS records. Over the next two years, assuming DLPWG's plans for centralization of digital assets into DSpace proceed well, an-

Libraries are using METS to record both presentation and preservation information so that digital assets can be accessed and presented both now and in the future.

other 50,000 or more METS objects for assets created at SIO, MSCL, and the Music Library will be added to UCSD's overall METS collection. The magnitude of this corpus will, I believe, lead the library to experiment with developing effective, automated preservation and management routines for its digital assets.

TR: Since you mentioned schemas and profiles can you differentiate between a METS schema, a METS document, and a METS profile?

BW: Sure. A schema defines the rules for creating METS documents. It specifies the data elements, the kinds of values that are permissible, and the relation among the elements. A METS profile is a set of rules for constraining application of the schema

for a certain class of materials, for example, electronic theses and dissertations, in order to assure interoperability among members of the document class. A METS document is an application of the schema to describe a single instance of an asset and its metadata.

As a crude analogy, MARC is generally the equivalent of the METS schema; MARC formats are, like profiles, for classes of materials; and MARC records are the equivalent of METS documents.

TR: What are the major sections of a METS document and can you explain them in laymen's terms?

BW: There are seven possible sections to a METS document: METS header, descriptive metadata, administrative metadata, file inventory, structural map, structural links, and the behaviors section. The METS header is where information about the creation of the METS document is stored. This could include the file name for the METS document, the date it was created and modified, the name of the person responsible for creating the document, etc.

The descriptive section is the equivalent of the MARC record in that it is for metadata supporting identification and interpretation of the digital asset as an intellectual construct. The descriptive information can either be stored (embedded) in the METS document, using MODS or Dublin Core, or it can be externally referenced, or linked to, as, for example, a MARC record in ROGER.

The administrative section provides information describing technical characteristics of the assets, their analog sources when they exist, and how the digital versions were created. Administrative metadata can also include rights information and descriptions of events that resulted in a modifica-

(continued on page 3)

METS (continued from page 2)

tion of the digital asset, such as migrating an asset forward from one file format version to a newer file format version or changing the original content in some fashion, say for example, adding a fourteenth person to the table in Da Vinci's *Last Supper*.

The file section lists all digital files referenced by the METS document. For a simple image, that might include an archival TIFF, a user JPEG, and a reference thumbnail GIF. For a multi-part object, it could consist of many files, each file having its own technical metadata. The file section is simply an inventory.

The structural map is the basic rationale for the METS standard and is where the relationship of the files listed in the file section is expressed. It provides the means for developing tools to allow easier end user navigation of complex digital assets, that is to move through files as though they were pages in a book (physical structure), or entries in a diary (logical structure), or both, say a scrapbook filed with a teenager's memorabilia. The structural map also provides the means for associating each file with its corresponding descriptive and administrative metadata.

The structural links section, with which I am least familiar, allows for recording hyperlinks across the divisions indicated in the structural map. As the METS documentation notes, this section is designed, in part, to support using METS for wrapping websites.

Finally, the behavior section associates executable behaviors with content in the METS document. For instance, it is possible to specify an application file needed to use the resource or to indicate the resource is to be displayed using a specific style sheet.

TR: Are there similar standards to METS and why did you feel METS was the best option for UCSD?

(continued on page 4)

Digital Library Program: Communication Survey Results

The Digital Communications subgroup of the Digital Library Program Working Group (DLPWG) sent out a survey to all library staff in January, asking what you think of our vehicles for communicating about the digital library program and related news. We'd like to share a summary of the results here, and also to invite you to take another look at *diglet* and *Digital Letters* if you're one of the people who haven't read them before.

Digital Letters (This is our print newsletter, also available in PDF) URL: <http://gort.ucsd.edu/dlpwg/dletters/dlindex.html>

Over 80% of respondents had heard of *Digital Letters*. We asked which kinds of articles you find the most interesting, and interviews, topic features and project updates got the highest ratings. A couple of things we're looking into based on your feedback include work on the layout to increase readability and whether there's an easy way to turn the online version into HTML.

diglet (This is our blog featuring news items about digital library activities) URL: <http://gort.ucsd.edu/mtdocs/diglet/>

We were somewhat sad to learn that less than 40% of respondents had heard of *diglet*. We hope you'll check it out - a dedicated team of bloggers led by James

Jacobs updates *diglet* almost daily on topics including privacy, open access, copyright, metadata, interesting readings and events.

Digital Dialogs (This is our series of lectures, webcasts, and brown-bag discussions, co-sponsored by the LAUC-SD Research & Professional Development Committee) URL: <http://gort.ucsd.edu/dlpwg/dialogs.html>

While many respondents hadn't had a chance to attend one, over 85% of those who did found the sessions to be useful. *Digital Dialogs* are announced to everyone, so keep an eye out.

We asked you to rate your interest in various topics that we've covered, and the top five, in order, were:

- UC digital library projects
- copyright issues
- public services
- privacy issues
- cataloging/metadata

We're hoping to use your feedback to make our digital library communications more relevant and more interesting, and we hope you'll continue to read them and participate in *Digital Dialogs*. If you have any comments, suggestions or questions, please contact Trish Rose, Chair of the Digital Communications Group.

~Jennifer Reiswig, Electronic Services Librarian at the UCSD Biomedical Library

METS (continued from page 3)

BW: A survey published by NASA in 2001 identifies several complex object technologies (<http://techreports.larc.nasa.gov/ltrs/PDF/2001/tm/NASA-2001-tm211426.pdf>). Of those reviewed in the survey, FEDORA, METS, and MPEG21 have most captured the attention of digital libraries. Of these three, METS is probably the most mature, albeit it is still very young and emerging as a standard.

METS has several virtues that make it a good option for the UCSD Libraries to adopt. First, it is well rationalized and intended to support management functions and interoperability. It is often referred to as an important part of any real world implementation of the OAIS reference model. Second, it is a standard born in the library community, having been developed initially to support the LoC's "Making of America" project at the end of the 1990s. The more libraries adopt it, the more likely it is the standard will become more robust and less costly to apply. Finally, it is the standard elected for use by the CDL and its Digital Preservation Repository, with whom the UCSD Libraries will collaborate in preserving the assets of the UCSD Libraries.

TR: How are METS documents for UCSD objects created?

BW: UCSD Libraries is currently engaged with batch production or assembly of METS documents for legacy materials, that is digital assets already created and already having some form of metadata. These materials have been created over a period of years by different library departments and some of the materials are at risk because of the media on which they are currently captured, for instance, assets stored on CDs. The aim is to centralize all these assets in one digital repository in a manner that facilitates efficient management and, at some point in the future, access by end users.

There are several steps to this process. First we identify the assets - where they are, how many there are. Second we identify what types of metadata exists for them - descriptive, technical and rights. If necessary we construct a data map for transforming extant metadata to types of metadata used in the METS document. For example, in the case of the AAL digital images, MARC records will be transformed into MODS records.

We then create what we call a decomposition exemplar, which describes a fully built METS document for a class of materials and indicates the source of each and every data value in the exemplar. The exemplar is delivered to an ETL team member in the IT Department. ETL stands for extraction, transformation and loading. ETL staff use the exemplar as a target for formulating procedures and code for assembling the METS documents. We've tested this

successfully for a small group of 50 or so METS documents, and, once other contingent problems are resolved, we will be ready to assemble METS documents for all the AAL digital images and for about 50 Film and Video Library objects.

At some point in the future, we will need to develop tools to support the production of METS documents at the point of first acquiring or creating assets and metadata. This would essentially transfer responsibility for the production of METS documents from the ETL team to the department staff responsible for the assets.

TR: What are some challenges with regards to METS?

BW: Great question! Three problems attract my attention: interoperability of profiles, expressing structure, and utilizing the metadata in METS documents to build automated and effective procedures for managing the library's digital assets.

Profile interoperability has to do with how well profiles work across institutions. The challenge is to guide profile development in a manner that supports sharing METS documents across institutional boundaries. The danger is that profile development could evolve in such a manner that intra-institutional interoperability is fostered, but at the cost of inter-institutional interoperability. It's something to keep in mind.

As for structure, it can be expressed at very gross or very granular levels. For instance, in regard to a musical recording, you can simply express structure at the level of a work or package such as a CD or LP, or you could express it at finer levels of movements or even themes. While those finer levels provide considerable benefits to end users, they also require manual indexing, which is never error free or cheap. It will be interesting to see if the increasing magnitude of our digital collections requires more and more granular structural expressions.

One of the great potential benefits of METS technology is the ability to develop automated processes for managing a digital asset collection. We would like to develop processes automated to the fullest degree possible, thereby minimizing the need for more expensive manual intervention. But it's very early in the day, and we still do not have a good idea of what metadata we need for many of the management processes we can imagine. It is hoped that the work of PREMIS, the group sponsored by RLG & OCLC, will expose some valuable strategies (<http://www.oclc.org/research/projects/pmwg/>).

TR: Thanks Brad. This discussion was extremely enlightening and gives us a better sense of why METS is important both within the library community as well as at UCSD.