Report on The Making of America II DTD
Digital Library Federation Workshop

Jerome McDonough, Leslie Myrick & Eric Stedfeld
New York University Libraries

Attendees

Columbia University:	David Millman
Cornell University:	Peter Hirtle, Sandy Payette
Harvard University:	Mackenzie Smith†, Robin Wendler
Library of Congress:	Morgan Cundiff, Carl Fleischhauer, Sally McCallum, Dick Thaxter
MetaE:	Alexander Egger
Michigan State University:	Mark Kornbluh
National Archives & Records Administration:	Dan Jansen, Kate Theimer
New York Public Library:	Terry Catapano, Bob Decandido
New York University:	Peter Brantley, Jerome McDonough†, Leslie Myrick, Eric Stedfeld, Jennifer Vinopal
San Diego Supercomputing Center:	Reagan Moore
University of California:	Howard Batchelor, Rick Beaubien, Bernie Hurley, Merrilee Proffitt†
University of Chicago	Charles Blair
University of Virginia:	Worthy Martin, Daniel McShane, Daniel Pitti

† Workshop Coordinators

Thursday, February 15th

Introduction

The workshop opened with comments by J. McDonough (NYU) regarding the origins and purpose of the workshop. University research libraries that have begun to create digital collections have encountered some common problems surrounding metadata and encoding for digitized versions of primary source material, including:

- inadequacy of traditional descriptive metadata schemes for describing digital objects;
- the absence of standards for administrative metadata; and
- the need for a flexible mechanism to express structural metadata regarding digitized versions of a wide range of materials in different formats (text, image, audio, video, etc.).

The Making of America II project (MOA2), sponsored in part by the Digital Library Federation, attempted to address some of these issues, and created an XML DTD based on the research the project participants had done on descriptive, administrative and structural metadata schemes. This XML format provided a single document type that could encode descriptive, administrative and structural metadata for a wide range of materials. However, it was intentionally restricted to textual and image materials, and so does not adequately support the needs of those trying to create digital libraries of audio-visual materials. A variety of other shortcomings of the DTD have also been identified by those trying to employ the DTD. This workshop was intended to try to examine the use of MOA2 to date and see if a successor format could be created which would rectify the MOA2 DTD's shortcomings.

Reports from the Field

The MOA2 DTD was used for encoding of digital objects in the original MOA2 project, and was subsequently modified and adopted as a University of California (UC) system-wide standard by the California Digital Library. The Library of Congress has also considered using the DTD, and Harvard University has employed it on some of its digital library projects. Staff from UC Berkeley, Library of Congress, and Harvard presented their experiences with the DTD to date.

M. Proffitt from UC Berkeley described both the MOA2 project's experience and UC's experience with use of the DTD. UC has added several elements to the DTD since the original MOA2 project in order to support new endeavors. Most of these additions have been in the form of new descriptive metadata elements, and modifications to the structural metadata sections of the DTD. The California Digital Library and the Museum and Online Archive of California are currently using the DTD, as are some internal projects at UC Berkeley. Use of the DTD has revealed several weaknesses, most of which have been addressed by the addition of new elements to the DTD. Use of the DTD has also required ongoing educational efforts as to best practices for using the DTD.

Discussion of the MOA2 project among the participants highlighted the fact that some of the tools created by the MOA2 project to assist in creating MOA2 objects did not work well in

environments outside UC Berkeley. These difficulties stemmed from the fact that the tools made assumptions about work flow and practices that were not accurate in some organizations participating in the project. Some of these problems led participants to leave out certain metadata from their MOA2 objects, particularly technical metadata. Use of the DTD on the original project also raised several unresolved issues regarding what exactly a digital object was, and how descriptive metadata should be used in conjunction with the DTD.

C. Fleischhauer from Library of Congress (LC) reported on LC's deliberations regarding the MOA2 DTD. Four issues led LC to consider using MOA2 for some of their projects: 1. the somewhat ad hoc nature of metadata created for the American Memory Project (especially non-descriptive metadata); 2. MOA2's potential applicability to encoding audio-visual materials; 3. LC's need to engage in large scale reformatting of taped material; and 4. the emerging issues and requirements for preserving content in digital form. Dick Thaxter discussed LC's effort to modify the DTD for their own using an outside contractor.

M. Smith presented Harvard's approach to using the MOA2 DTD. Harvard did not encode administrative information within their MOA2 objects, preferring to store this information in a database. MOA2 serves primarily as a structural metadata system for their projects. Harvard's use of the DTD has differed somewhat from UC/CDL's, as Harvard has used a large number of MOA2 files to represent a single 'object' structurally, rather than trying to include all the structural metadata in a single file. Since an 'object' in their case might constitute a journal run, use of multiple, linked MOA2 files simplifies encoding, and improves performance for displaying objects.

Related Standards

After a break, discussion turned to consideration of a number of other existing and emerging standards that might influence the design of a successor format to MOA2. Standards discussed included Resource Description Framework (RDF), Synchronized Multimedia Integration Language (SMIL), MPEG-7, NISO Technical Metadata for Digital Still Images, the National Library of Australia's Preservation Metadata for Digital Collections, the Interoperability of Data in E-Commerce Systems (INDECS) proposal, and descriptive metadata standards including Dublin Core, MARC, Encoded Archival Description (EAD), and the Visual Resources Association Core Categories.

The discussion began with an examination of standards that support encoding of structural metadata, including RDF, SMIL, and MPEG-7. The discussion regarding RDF focused primarily on whether a successor format for MOA2 should use RDF, with RDF Schema as the mechanism for defining the format, rather than XML (via an XML DTD or Schema).

Discussion then turned to SMIL and MPEG-7. SMIL provides an XML format for encoding the structure of multimedia works, along with additional metadata. SMIL could be used as a replacement for MOA2, although it does not have any defined descriptive or administrative metadata elements. SMIL is the product of an industry consortium, and not a true standard. MPEG-7, by contrast, is being developed as an ISO standard and should be completed this year.

Structurally, MPEG-7 is far more focused on audio-visual materials than SMIL, and is not intended to handle the range of materials that SMIL was designed to accommodate.

Administrative metadata standards discussed include the NISO Technical Metadata for Still Images, the National Library of Australia preservation metadata set, the Library of Congress Audio Visual metadata, and INDECS. While technical metadata sets such as NISO's and Library of Congress's have become quite sophisticated and a consensus seems to have formed about their structure and contents, metadata structures for rights management and preservation purposes still seem inchoate. While there is interesting work being done by CEDARS and NedLib, preservation and rights metadata do not seem to have achieved the level of consensus that has developed around technical metadata standards.

Discussion of descriptive metadata included Dublin Core, MARC, EAD and VRA Core. Both Dublin Core and MARC have been translated into markup language encodings that could be integrated into a MOA2 document format. There was a great deal of discussion about the relationship between EAD and MOA2, and where the boundary between the two document classes should be drawn.

Friday, February 16th

Requirements for a MOA3 DTD

Having reviewed relevant standards, the discussion then turned to what participants might want to see in a successor format to MOA2. C. Fleischhauer indicated that Library of Congress needs a format that supports the variety of work in which they are engaged, including reformatting projects. A successor format would need to accommodate a wider range of metadata regarding source physical objects, treatment of those objects, digitization workflow and technical metadata for digital material. M. Smith talked about Harvard's needs in a successor format, including introducing greater flexibility into the format by eliminating as much as possible the use of required elements, and adding support for disseminating a wider range of materials, including audio-visual materials.

There was a great deal of discussion regarding MOA2 and any successor's relationship to the Open Archival Information System (OAIS) reference model. Harvard would like a successor to serve primarily in the role of a Dissemination Information Package (DIP), incorporating structural metadata and descriptive metadata that might be of use to patrons. In the original MOA2 project, the MOA2 documents served more of a Submission Information Package (SIP) role, and the University of California continues that practice, with MOA2 providing the format for documents being submitted to a central repository. The possibility of MOA3 serving as an Archival Information Package (AIP) was also discussed. While some saw potential value to this, others felt that the contents of an AIP needed to be determined by local needs and practice, and could not be determined by a community standard. It was also felt that formulating an AIP would require mandating administrative metadata elements when such sets are still in early stages of development.

A consensus emerged from these discussions that for a successor format to be useful, it must allow for the use of multiple descriptive, administrative and technical metadata schemes. Additionally, as much of the document format should be optional as practical, to allow for flexibility in local use of the format. A single standard for structural metadata, however, was seen as one of the benefits of MOA2, and there was agreement that this should be expanded to support audio-visual works. There was some further discussion of whether structural metadata would need to support geospatial data or other data formats. It was agreed that while this might be desirable in the future, a format that supported text, still images, audio and video would fulfill most libraries immediate needs for a document encoding standard.

There was also discussion of an appropriate document encoding specification mechanism, and whether a successor format should employ RDF or XML, and if XML, whether to use DTDs or a Schema. The consensus of the group appeared to be that XML would be preferable to RDF, and that Schemas probably provide a better mechanism for defining the document structure, as long as suitable tools for validating documents against a Schema seem to exist.

Discussion then turned to SMIL and the degree to which structural metadata in the document format should draw upon SMIL's modules. While SMIL was seen as interesting, the lack of development tools for creating SMIL material was seen as problematic. There did not appear to be strong interest in basing a successor format on SMIL. However, it was also felt that if a successor format could be constructed in such a way that automated conversion into SMIL was possible, this might allow libraries to take advantage of SMIL tools if they become more widely available.

The exact requirements for structural metadata were discussed. The hierarchical structure employed by MOA2 does not appear to be a problem, but its facilities for identifying/pointing to portions of a file/stream were seen as inadequate. There was consensus that a new format should be able to identify a particular point, a range (e.g., a segment of an audio file or a text file), or a block (e.g., a two-dimensional portion of a still image or video segment) as appropriate in any of the four major classes of data file (text, images, audio, video).

Next Steps

Attention then turned to the issue of who might develop and maintain a new format. J. McDonough volunteered New York University's digital library team to develop the schema. NYU will try to develop and distribute a draft of this schema in time for discussion at the DLF Forum in May, 2001. B. Hurley said that U.C. Berkeley is willing to be a test bed for documents, if there are people who will create objects to send to U.C. R. Beaubien and M. Smith volunteered to provide consultation and support to NYU's team in developing a schema.

There was general agreement that tools for creation and display of documents were crucial to the format's success. Tools need to be adaptable to local needs. Equally, however, no one wants to recreate the wheel. It was agreed that a crucial role for a maintenance agency would be to serve as a clearinghouse for knowledge on tool development, use, etc.

This lead to a discussion of who would serve in the role of maintenance agency. B. Hurley indicated that U.C. Berkeley was willing to help with tool development and testing, but that he was uncertain as to whether it was wise for U.C. to assume the role of maintenance agency. D. Pitti mentioned the two distinct roles played by Library of Congress (maintenance) and the Society of American Archivists (development) with respect to EAD. S. McCallum indicated that Library of Congress would be willing to assume a maintenance role with respect to a successor format similar to that which it plays for EAD, but does not have the resources to be the lead on future development work on the standard. The possibility of DLF assuming the developer role with respect to MOA2 was raised. J. McDonough said that he would discuss the issue with Dan Greenstein of the DLF and report back.

At this point, the formal meeting ended. An informal discussion continued on whether specific metadata sets for descriptive, administrative and technical metadata might be required or recommended. The consensus seemed to be that such a move was premature. This poses a potential problem for tool builders, since they will need to create tools flexible enough to cope with arbitrary metadata that may be embedded in a MOA3 document. J. McDonough proposed something similar to Z39.50 profiles, where particular combinations of metadata sets could be named and registered. This is obviously an area where further work will need to be done.