Metadata and Digital Projects: What Every Archivist Needs to Know

By Kristen Walker

2005

Metadata, otherwise known as information about information, has always been a constant in one form or another in the archival environment no matter the international boundaries. Today there is an increasing amount of metadata that archivists need to be aware of thanks in large part to the multitude of digital projects going on around the world. Some of these metadata standards archivist should at least have a working knowledge of include the emerging Encoded Archival Description standard, the Encoded Archival Context standard, and several others including TEI (Text Encoding Initiative,) MODS (Metadata Object Description Schema) and METS (Metadata Encoding and Transmission Standard.) Archivists can no longer afford to hide behind their proverbial stacks of paper, but need instead to embrace these new standards, take a proactive stance and pursue these standards to the best of their intellectual ability through whatever means is available to them. The more archivists know about these standards the more comfortable they will be in the digital environment.

EAD

As an archivist I have had a vast amount of experience with a variety of metadata formats. One of the first standards I learned post graduate school was the Encoded Archival Description. During the summer of 2001 it was decided by the then Dean of the Indiana University Libraries that the Lilly Library and the Indiana University Archives needed to learn and implement EAD in their repositories. That fall, during the Midwest Archives Conference meeting in Indianapolis, Indiana, I had the opportunity to take part in The Society of American Archivist's (SAA) workshop on EAD taught by Jackie

Dooley and Richard Szary. These lessons were brought back to the archives and shortly thereafter an EAD working group was formed by the then Associate Dean of the Libraries. The School of Library and Information Science (SLIS) also developed and funded an EAD graduate assistant position during this time. This person worked at both repositories encoding our finding aids while also helping further develop and establish the archives specialization in the library sciences program.

After some growing pains and a vast amount of reading on the topic my colleague, Dina Kellams, and I became well versed in the practice of EAD and were approached by SLIS faculty to lecture to their classes on the subject. We were also offered the opportunity to teach a 1.5 credit workshop on EAD. The class has proven popular among our SLIS students and has been offered the last two summers. There are two important lessons to take from this example. The first is archivists taking the initiative to explore and learn everything they could about this metadata standard so that it was applied correctly and so we could teach others to do it as well. It is also important to note that this educational process never ends. It is a constant battle keeping up with all of the literature coming out on this subject and to be knowledgeable about the accompanying standards such as the ISAD(G)[1]. Most recently I took part in the EAD to MARC web based workshop offered by the Society of American Archivists (SAA) which was taught by Michael Fox, I will also be taking the *DACS-Describing Archives a Content Standard* workshop offered by SAA later this month at the Midwest Archives meeting in Bloomington, Indiana.

[1] International Council on Archives. *General International Standard Archival Description*, Second Edition, 2002, http://www.ica.org/biblio/cds/isad_g_2e.pdf

The second lesson was by learning EAD I became curious about other metadata standards and how they could be applied to other applications at the University Archives. This led me to pursue and learn TEI (The Text Encoding Initiative) on my own and to purse readings on other metadata standards. I have also pursued closer working relationships with my friends in the Digital Library program and have found them to be an invaluable source of information.

Metadata and Digital Projects in the Indiana University Archives:

In the spring of 2001 the Indiana University Archives began work on the digitization of the Indiana University Board of Trustee Minutes. Initially the project started out with the simple vision of making all of the Boards minutes available via the web to our patrons twenty four hours a day, seven days a week thus allowing board members, university counsel, and other university administrators access to critical information contained in the minutes on university policy and action outside of the archives typical hours of operation. With this simplistic vision came the ultimate realization of a much more complex and at times daunting project which not only has challenged the Archives and Board of Trustees office staff, but also the staff of the Indiana University Digital Library as well.

Prior to the archives work on this digitization project the minutes were only available in paper format. Researchers would have to trek over to the archives during limited hours of operation and fit their research into busy schedules. Often times administrators, faculty, and students found themselves to be too busy to do this and all research would be left up to either the staff at the archives to complete or to the staff at the Board of Trustees office to do.

Due to the lack of access points, requests for material contained in the minutes could take up vast amounts of staff time to research and locate. This resulted in an impact in the amount of work the staff was able to complete in a day when these requests came in. It was conceptualized that if the minutes were available and searchable online administrators and researchers would be empowered to do their own research and would also allow the archives staff to concentrate on other pressing needs.

There was also the issue of preservation of the bound volumes containing the minutes to consider. The minutes in this collection run from 1884 through the present with one odd volume from 24 September 1838-15 July 1859. The rest of the minutes dating back to the founding of the university in 1820 and including the American Civil War era were lost in the great campus fire of 1883 which burned the entire campus down to the ground.[2] By digitizing and making available via the web the minutes of the board we hope to be able to effectively limit the amount of handling these bound volumes undergo, possibly preventing new damage to them.

At the outset of the project, funded by the University Board of Trustees Office it was decided to encode the minutes in reverse chronological order since our main constituents needed access to the most recent minutes first. Having adopted HTML as our initial format and quickly realizing its lack of functionality, we opted to start encoding the minutes using the Text Encoding Initiative, or TEI metadata standard.

The TEI standard was a metadata standard that the author was completely unaware of and ignorant to until I was introduced to it by the Indiana University Digital Library Program. Once examined it was determined that the TEI was not too unlike EAD

[2] Wylie, Theophilus A., *Indiana University, Its History from 1820, When Founded, to 1890, with Biographical Sketches of Its Presidents, Professors and a List of Its Students from 1820 to 1887*, Wm. B. Burford, Printer and Binder, 1890, preface.

in its structure and would be easy to learn. I studied the TEI website and available literature and in a short amount of time was able to come up with a template for my project. I based my template on the elements available to me through the TEI Lite DTD. So again, in this case, the archivist took a proactive approach to a new metadata standard. Not only did I need to learn it so a functional standard could be applied to my project, but I also needed to learn it so a template could be developed, and so I could communicate with my colleagues in the Digital Library Program about the needs and wants of the project in a coherent fashion.

In conjunction with encoding the born digital University Board of Trustee minutes the conversion of the minutes available in paper format only from 1970 through 1985 into electronic text files was started. Scanning was done using ABBYY software recommended to us by our contact at the Indiana University Digital Library Program. All of the minutes were scanned into TIFF files at 600 dpi and were compressed to the Group 4 setting to minimize the file sizes. Once scanning was completed for a calendar year the TIFF files were duplicated to Word documents for encoding and the TIFF files were zipped together and put into storage in the University's HPSS (High Performance Storage System.)[3] Once the student staff had completed the process of duplicating the TIFF files to word files they were meticulously proof read and compared to the original minutes. So far the margin of error has been acceptable and we have not yet had to go to a double keying strategy. However, we will need to do so for the handwritten and some of the typewritten text.

[3]Indiana University Knowledge Base. "What is HPSS?" (July 20, 2005) http://kb.indiana.edu/data/agvi.html?cust=087117.79308.131 (Accessed July 25, 2005)

As of the end of July 2005, five years worth of minutes for the University Board of Trustees, approximately 50 set of minutes with an average of 30 pages per set, have been encoded. A website where these minutes will be published is currently under development. Each set of minutes will be published using the XTF software package currently in use by the Online Archive of California[4]. As a side note XTF is also being explored as a possible delivery system for the finding aids created at the university as well and may replace the DLXS (Digital Library eXtension Service) [5] package currently in use at the university. There are also plans in place to link these published minutes in the accompanying EAD finding aid so no matter how the collection is accessed the user will be able to navigate to the full text of the minutes. There have also been discussions of developing a user study, but talks on this are only in the very early stages.

To aid in the creation of access points to the digital minutes the individuals involved with the project interviewed patrons who use these materials the most to see if we could customize our encoding practices to reflect their needs. In response all names, campus buildings, and dates were regularized in the attributes of the elements used in the text to accommodate the requests we received. Often administrators or faculty want to be able to look themselves up to see when promotions were officially made or what they said at a particular meeting for example. We opted to regularize these names based on the format used for authority records at the Library of Congress. Often times we were able to locate faculty on the Library of Congress Name Authority website[6] since as part

---

[4] California Digital Library. "Inside CDL: eXtensible Text Framework (XTF)." (Dec. 14, 2004) http://www.cdlib.org/inside/projects/xtf/ (Accessed July 15, 2005).
[5] University of Michigan. "DLXS: Digital Library eXtension Service." (Dec. 12, 2003) http://www.dlxs.org (Accessed July 15, 2005).
[6] Library of Congress. "Library of Congress Authorities." (April 26, 2005) http://authorities.loc.gov/ (Accessed July 12, 2005).

of their tenure process they had produced published works. For unpublished individuals we collected as much information as we could, often from the Faculty Records office and created an Excel spreadsheet containing a list of local authority files. We did this though a partnership with our technical services department who may want to use some of our local authority files. We will also be using our local authority files for our photographs database which is currently under development.

Future Metadata Possibilities:

In the future we may look into adopting the EAC (Encoded Archival Context) into our daily workflow. The EAC is a standard structure for the recording and exchange of information about the creators of archival materials.[7] After attending the International Congress on Archives in Vienna, Austria, in 2004 we tentatively did some research on the subject. We are eager to hear more about the testing of the beta version and also to receive our training with the *Describing Archives: A Content Standard* before we pursue it too far so as to become more familiar with the ISAAR (CPF) standard and how it is applied here in the United States. However, I would like to stress that we are interested in it and will monitor its future progress.

Finally, some other metadata standards that archivist working with digital projects should be aware of include the METS (Metadata Encoding & Transmission Standard) and MODS (Metadata Object Description Schema.) As a brief description, METS is intended to document the structural and technical metadata regarding the metadata process so that the integrity of all the digital images can be maintained. METS is more

[7] Yale University. "Encoded Archival Context (EAC)." (March 3, 2003) http://www.library.yale.edu/eac/ (Accessed August 1, 2005).

extensive and different in function than MARC21.[8] At the Indiana University Archives METS is being used in conjunction with another digitization project that is currently ongoing involving the records and papers of Andrew Wylie, the first president of Indiana University, and his family.

MODS is "an XML schema which is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format. This schema is currently in draft status." [9] It can be used as an extension to METS and is basically meant as a complimentary piece to other metadata standards. As another standard in draft the Indiana University Archives is not currently using this standard, but are eagerly awaiting results of testing being conducted by the Indiana University Digital Library program to see if this is a standard we are interested in adopting into our digital project workflow as well.

Conclusion

In concluding this paper I think that archivists need to take a more proactive approach to learning all of the metadata standards that are out there and to have a working knowledge of the international standards behind them. I would also argue that more education at the graduate level for archivists on managing digital projects and becoming proficient with all of these metadata standards should be stressed. As for archivists already practicing perhaps a workshop formed in partnership with the Society

---

[8] The Library of Congress. "METS: An Overview and Tutorial," (May 24, 2005) http://www.loc.gov/standards/mets/METSOverview.v2.html (Accessed August 1, 2005).
[9] The Library of Congress. "MODS: Uses and Features." (April 7, 2003.) http://www.loc.gov/standards/mods/mods-overview.html (Accessed Aug. 1 2005).

of American Archivist and the Digital Library Federation could be explored for example? There are many more things we can do as individuals to educate ourselves about metadata and all of the useful applications it can have in our archives. I hope some of you will be inspired by this paper and the projects described in this session to look into them further and to apply them in your institutions to the best of your personal ability!