

Amy M Ostrom  
LIS651 – Cataloging & Classification  
7 May 2004

## Digitizing with the Digital Content Group

With the advent of digitization technology, many organizations are turning from their print sources to electronic facsimiles for mass distribution and access. Not only the hardware, such as computers, large scanners, and storage devices, but software that is being created enhances yearly for a more unified and formal approach to producing and sharing information online. Among the most important software being developed are the metadata and encoding languages that provide structure for documents. Encoding languages are essentially the rules to how metadata can be presented in the document, using tags to define each type of metadata. Following the brief explanation of metadata and history of encoding languages and their structure, I will give an in-depth example of the process needed to digitize an item.

Metadata is described as data about data. With an online document, metadata is typically divided into specific elements that are designated by tags. Some organizations will break metadata into several different types, but for most purposes three overarching types exist: administrative, structural, and descriptive. Administrative metadata is typically found at the beginning of each document and contains information about what organization created the document, when, with what tools, with whose permission, etc. Normally this information is not visible to the end user, but is used more as liability and for internal purposes. Structural metadata is used for page layout, essentially telling the information where to show visually. An example of this would be something such as the <table> tag in HTML. It is also the metadata that is used to define containers of elements and the order in which they appear (not visually, but in constructing the document), an important feature in XML. The last type of metadata is the descriptive kind that is used to actually describe what the information is. The <address> tag also in HTML is a good example of this, in which one will put the contact information of the person who built the document.

SGML, which stands for Standard Generalized Mark-up Language, is one of the earliest known standard encoding languages that came out of research done by the American National Standards Institute (ANSI) and Charles Goldfarb of IBM. SGML's encoding structure does not allow for Internet display of each element without a separate document to prescribe visual effects to each element used in the document. The language made its first appearance in 1980 with a purpose to provide a set of rules to structure online documents. This became the goal of encoding languages with the use of an item known as a tag. SGML passed this trait down into its sibling languages, HTML and XML.

HTML and XML stemmed from SGML as needs for a simpler and smaller set of the tags became necessary with the Internet's growing popularity. HTML, or Hyper-Text Mark-up Language, is based on a set of predefined tags inherited by SGML. Tim Berners-Lee proposed this highly simplified set specifically for the displaying of information on the Internet. These tags do not allow for defining the type of information that is contained inside of them, but rather to define the layout and visual appearance of the information. For example, the <b> tag will produce bold text inside of its tags.

HTML did not fulfill the need of structuring a document in an understandable sequence, and so XML was also created outside of the original SGML set.

XML, or Extensible Mark-up Language, was built because of HTML's lack of ability to define individual pieces of information in a logical structure. XML does not have predefined tags such as HTML (except for a handful of predefined base tags that are needed for XML). XML tags are created by the user to represent the information in an understandable structure. For instance, if one was to build an online document to represent a directory, some tags that could be created would be <entry>, <name>, <company\_name>, <street\_address>, <telephone>, etc. For the browser to be able to validate the document's tags, it needs a DTD, or document type definition. XML requires that its structure be "well formed," or that tags contain information and are within the correct container tags according to the DTD that is referenced. XML, like SGML, then requires a stylesheet to be attached to display the information in a visually appealing manner (without the tags). Although XML only contains a small subset of SGML's abilities, both SGML and XML require that a DTD be assigned to the document.

DTDs are vital to XML, since they define what each tag is. Many DTDs are already defined and are available to use through several online repositories such as at [www.w3c.org](http://www.w3c.org). Some of the most well-known DTDs in library science are TEI, or Text Encoding Initiative (see appendix for sample), and EAD, or Encoding Archival Documents. The DTDs that have been submitted and made available at a site such as [www.w3c.org](http://www.w3c.org) are also known as standards. These documents contain several tags that were created by committees and are constantly revised to reflect the growing needs of digitized information. TEI has undergone several changes itself and is on its fourth revision. It has also created a subset DTD of some of the most commonly used tags of the TEI set of about 700 elements. This subset of TEI is known as TEI-Lite and has roughly around 100 of the original TEI element tags. Once an individual decides to use a DTD that has already been defined, that person must use the tags properly according to the DTD and documents provided by the group or individual that built that DTD. DTDs may also be expanded if the original does not contain the required elements. Some do not try to conform to another group's set of tags and thus creates their own DTD. DTDs do not need to be validated by any group for them to work, but the document using the DTD must be validated to the DTD. The University of Wisconsin-Madison has chosen to use the TEI DTD with great emphasis on the TEI-Lite tags in TEI, since it does cover the majority of the elements that can be found in a book.

The University of Wisconsin-Madison has become involved in digitizing books accessible on the Internet. The group assigned to do this is known as the DCG, or Digital Content Group located on the fourth floor of Memorial Library. It consists of less than a dozen full time staff and another dozen or so student workers that work together in its small one room office. Each of the staff members controls different steps of the process of digitizing materials, from its initial request to putting its presence up on their website. Because the DCG does not get too much funding for general collection digitization, requests must be placed by relevant members of the University for any material to be considered for digitization.

One particular project is known as the Digital Library for the Decorative Arts and Material Culture. This project is being funded by an organization unaffiliated with the University known as the Chipstone Foundation. The Chipstone Foundation was founded in 1983 by Stanley and Polly Stone for the purpose of preserving their collection of

decorative arts from 17<sup>th</sup> and 18<sup>th</sup> century America. Creating the online digital collection is just one of the many ways the Foundation is actively involved in keeping the decorative arts an interest in the Wisconsin community. Contacts from the Chipstone Foundation have allowed for the DCG's Linda Dychak to choose materials out of the University's collection to digitize through their funding that fits the criteria laid out by the Stones.

A typical project that is given to the DCG has papers that are filled out by the requesting party. Because the Chipstone Foundation has good faith in the DCG's selections, it is the DCG that puts together the assessment documents for this particular project. All other projects have the requestees fill out the assessment papers required to determine whether a certain material is proper to digitize or not. Many aspects are considered in this selection process. One of the most important questions asked in the beginning is whether the digitizing will be a facsimile or simply an electronic version of the material. The difference in these two lead to very strict outcomes. The e-facsimile collection the DCG has produced contains materials that have been completely digitized. This includes the covers, the blank pages, anything that does not carry direct information, and all pages that do have information. The Decorative Arts and Material Culture project is an e-facsimile project, so all material chosen is scanned literally from cover to cover.

In the project assessment, the questions asked are needed in order to verify the item is a good candidate for the digitization process. The project is given a name, then questions about whether the item has been digitized by another organization and if so what format (microform, book, etc) has been used, accessibility to the item, and purpose for digitizing are some of the most important requirements. Another interesting question that is brought up is whether a book may be disbound for speed scanning. This is a very serious question for the importance of preserving the book in its original form. Some potential candidates are borrowed from outside of the library, and it is important to know whether the DCG may remove the binding or not. Once these questions are filled in, members of the DCG sit down and review the request. If a material is thought to be a good candidate but has substantial issues, the item may not be digitized due to lack of technology.

The project I worked on while with the DCG was a book from 1764 containing over fifty plates of the emperor Diocletian's palace in Spalatro. The book was in relatively good condition with minor spinal damage. It was a very peculiar book due to its size however, and we had to determine whether it was a good candidate to put the effort into digitization. We also found the microform of the book had been digitized, but the quality of the images was far inferior to what we could produce with the book, so we concluded it would be digitized with the new global scanner the University had just purchased that has the capabilities of scanning in color items that were three feet wide. This is necessary for my project.

Once it is decided that the material would be digitized, the DCG begins working with the material by creating metadata in an Excel document. This data is double-checked in what is called "quality control" in order to come to an agreement as to how the book will be represented online. After this occurs, the Excel sheets are sent through a program created by DCG staff that essentially strips away Excel's coding, and pulls out the inputted information. This is then put into an SGML document that will be read by the Open Text Database to create HTML pages on the fly. The metadata input with the Excel document is an exhaustive process that is essential in providing an easily accessible and functional online document. If one cell is missing and it is sent through the program,

the resulting SGML will be incoherent. The tag set for the SGML document is TEI, the crossover between the Excel data fields and TEI tags created by Peter Gorham of the Library Technology Group (LTG). The LTG is the group that takes the tabular data created by stripping away the Excel coding and makes it into a functioning SGML document.

Once the metadata is complete, all of the scanning is done to make the facsimile of the material. If the book is able to be OCRed (optical character recognition), where a program “reads” the image to create a text document from the scanned pages, the metadata will have two links – one for the image and one for the text file. Three types of scanning are available, depending on the book. If the book is to be disbound (removing the spine), it can be sent through a high speed scanner. The DCG does not rebind the book once it has been fed through the scanner since much of the margin has to be removed in order to cut off the spine. If this is not the case, it will then be done on either a flatbed scanner or a global scanner. This takes much longer to do as settings will have to be made and each individual page taken care of manually. The file names are very important and must match up to what had already been created in the metadata, so once all the pages are scanned in and made the appropriate image type using Photoshop, all the documents are combined and put on the test server for a final quality check.

The final QC is very important to find pages that were scanned poorly, unreadable, or just not there (broken link). The final QC is also known as the “click through,” since student workers must click through each page of the online facsimile to check for metadata typos and broken links. These errors are recorded, fixed, and when all pages work in a collection, the collection can then go live.

What was described was one of the two types of projects the DCG works with. The other type of project is based on images, not texts, so it has a slightly different approach. They use Dublin Core, which is where cataloging really ties into the digitizing process. Dublin Core is another standard that “is a simple yet effective element set for describing a wide range of networked resources.” (Hillmann) Although the PageTurner model does require cataloging decisions in metadata display and requires knowledge of AACR2r rules, the use of Dublin Core requires finding subject headings using MeSH, AAT, LCSH, etc whenever possible, and using ISO standards for date and language information. Dublin Core is not something that is seen when the HTML page is displayed, but it is essential in identifying the project. It is administrative metadata found in the <head> tag using the <meta> tags. The DCG will use data appropriate to library standards unless the client specifies otherwise (this would happen for an archives project that has subject heading created specifically for that collection).

One can see that digitizing a book that conforms to a library’s standards is not a very quick and easy job. Many times several hard decisions must be made, such as defining the plates in the book I worked on as leafs so the extremely large blank pages did not have to be replicated for the facsimile. The work has to be analyzed to determine book parts, so the display breakdown on-line makes sense to the user. In some cases books must be manually marked up in the DCG, wherein the student must actually read the book to determine where tags belong. QC is extremely important because the project would be worthless if it was completely scanned in and the end user was unable to access any of it or could not find it because of typos. This holds true with cataloging in that if the cataloging record is different from that on the book, the book could be lost on a shelf

somewhere. Digitizing is a long process, and requires good library skills to make it work as if the digital facsimile were the book.

Cataloging and classification is a very similar process to creating retrievable online documents. They both require standards to conform in a manner that allows items to be easily retrieved in an environment that is easy to learn. Cataloging, and more specifically classification, creates a call number based on arbitrary information to “bind” books with similar topics together. In a sense this can be viewed as metadata, since we are assigning a subject to a book that will define it. Likewise, metadata online is used to describe documents, just on a more individual basis. It would be like defining every paragraph, sometimes each word or sentence in a book to its own subject. The DCG ensures the longer library presence’s standards by incorporating its rules with the rules of encoding languages to produce its online materials. What we are seeing is a merging of two structures with similar needs, because without metadata to describe an item, we would find searching for the item much harder.

## Bibliography

Chipstone Foundation. (2003). *Chipstone*. [ONLINE]. Accessed 3 April 2004. <http://www.chipstone.org/>

Cover Pages. (12 June 2002). *Standard Generalized Markup Language (SGML)*. [ONLINE]. Accessed 28 March 2004. <http://xml.coverpages.org/sgml.html>

Hillman, Diane. (26 August 2003). *Using Dublin Core*. [ONLINE]. Accessed 17 April 2004. <http://dublincore.org/documents/usageguide/>

SGML Users' Group. (11 June 1990). *A Brief History of the Development of SGML*. [ONLINE]. Accessed 27 March 2004. <http://xml.coverpages.org/sgmlhist0.html>

## Appendix

### TEI DTD sample

```
<!ENTITY % TEI.core.dtd PUBLIC "-//TEI P4//ELEMENTS Core Elements//EN"
'teicore2.dtd' >%TEI.core.dtd;
<!--Define the top-level element for this DTD-->
<!ELEMENT tsd %om.RO; ((tagDoc | entDoc | classDoc)+)>
<!ATTLIST tsd
    %a.global;
    TEIform CDATA 'tsd' >
<!--Define some additions for the phrase level tags-->
<!ELEMENT gi %om.RO; (#PCDATA)>
<!ATTLIST gi
    %a.global;
    tei (yes|no) "yes"
    TEIform CDATA 'gi' >
<!ELEMENT tag %om.RR; (#PCDATA)>
<!ATTLIST tag
    %a.global;
    TEI ( yes | no ) "yes"
    TEIform CDATA 'tag' >
<!ELEMENT att %om.RR; (#PCDATA)>
<!ATTLIST att
    %a.global;
    tei (yes|no) "yes"
    TEIform CDATA 'att' >
<!ELEMENT val %om.RO; (#PCDATA)>
<!ATTLIST val
    %a.global;
    TEIform CDATA 'val' >
```

### TEI DTD customization sample

```
<!DOCTYPE TEI.2 PUBLIC "-//TEI P4//DTD Main DTD Driver File//EN"
'tei2.dtd' [
<!ENTITY % TEI.prose      'INCLUDE' >
<!ENTITY % TEI.linking   'INCLUDE' >
<!ENTITY % TEI.analysis  'INCLUDE' >
<!ENTITY % TEI.figures   'INCLUDE' >
<!ENTITY % TEI.XML       'INCLUDE' >
<!ENTITY % TEI.extensions.ent SYSTEM 'teilitex.ent' >
<!ENTITY % TEI.extensions.dtd SYSTEM 'teilitex.dtd' >
]>
```

### Dublin Core sample

```
<head>
<title>Ecology and Natural Resources Collection</title>
<!-- Dublin Core metadata -->
<link rel='schem.DC'
href=""http://purl.org/metadata/Dublin_core_elements'>
<meta name='DC.Title' content="The Ecology and Natural Resources">
<meta name='DC.Date' scheme='ISO 8601' content="2003-03-21">
```