Computers and the Humanities **33:** 175–184, 1999.
© 1999 *Kluwer Academic Publishers. Printed in the Netherlands.*

An SGML/HTML Electronic Thesis and Dissertation Library

JANET ERICKSON and MATTHEW STOEFFLER
University of Michigan

The Electronic Thesis and Dissertation Project (ETD) was launched in 1987 at an Ann Arbor, MI, meeting arranged by UMI and attended by representatives of Virginia Polytechnic Institute and State University (VT), the University of Michigan (UM), SoftQuad, and ArborText. VT funded the development of a Document Type Definition (DTD) for dissertations and theses. The project continued at VT, with collaboration from the Coalition for Networked Information, the Council of Graduate Schools, and UMI, among others. Since 1994, many VT students have submitted their dissertations and theses in Adobe's Portable Document Format (PDF). As of January 1997, VT requires its students to submit their projects in electronic form rather than in paper. The long-term plan is to have them submitted in both PDF and SGML. VT's ETD Project is part of the Networked Digital Library of Theses and Dissertations, funded by a grant from the U.S. Dept. of Education. VT has been joined by 25 universities in supporting this effort.

The aim of my initial project was to describe a potential online library of dissertations and theses at the University of Michigan. The focus was on the SGML markup of sample dissertations using the TEI DTD and an HTML-based user interface for searching and retrieval. These latter elements will be developed as the body of electronic dissertations at UM grows. The electronic dissertations described here can be thought of as extensions of their print counterparts. Software, multimedia projects, and other natively electronic submissions are a different animal entirely and are not addressed in this project. The discussion of the initial project is followed by an update on the realized project at the University.

I acquired four dissertations to show the breadth of types that would need to be covered by the selected DTD. The first is the doctoral dissertation of Rebecca Price-Wilkin on the architectural history of a French church, prepared for the UM Department of Art History. Price-Wilkin's document was used as an example of an image-rich dissertation. The second is the doctoral dissertation of David Ruddy on the medieval travelogue *Mandeville's Travels*, completed for the UM Department of History. Ruddy's shows how historic texts can be represented in this model. A third dissertation is from Michele Tepper, and discusses culture in modernist literature. Tepper's dissertation was incomplete at the time, but was written for the UM English Language and Literature Department. It demonstrates the diversity that can be found within a dissertation, as each chapter can be thought of as an independent unit. Fourth, I acquired the dissertation of William Wheeler on global warming and agriculture from the Pennsylvania State University. This document contains many tables, graphs, and formulas, allowing me to speak to these important issues. It was used only for demonstration purposes and will not be included in further UM electronic dissertation projects.

Selection of a DTD

The first challenge to the project was selection of a Document Type Definition (DTD), or the rules by which markup would be applied to each document. The ETD project at VT had completed an initial DTD (the ETD DTD, later called ETD-ML or Electronic Theses and Dissertation Markup Language) for use on theses and dissertations. I initially elected to use this DTD, as it seemed simple enough that students with a passing knowledge of HTML markup could learn to use it with minor difficulty. Also, it had been designed with the material in mind, so there was some anticipation of suitability to the task. Lastly, I had assumed that the DTD was well used and perfected such that we could examine previous research and example ETD-ML dissertations. Unfortunately, VT provided a link to only one dissertation marked up with this DTD and the author has restricted its use to the VT campus only.

The ETD DTD, though in beta, was incomplete when my investigations began. A line break or ⟨BR⟩ was not available and there was no additional containment for the different parts of a ⟨HEAD⟩ such as a subtitle. Footnotes were referenced with a ⟨LINK⟩ element that includes a required IDREF; the corresponding ⟨FOOTNOTE⟩ element to contain the actual text of the note was not included in any content model in the DTD. Use of this DTD was also hampered by the complexity of the four dissertations selected for the project. One had several lists and an introduction that both preceded the chapters; the ETD DTD did not have structures for these, so the lists would have to go into the first ⟨CHAPTER⟩ element. The ⟨CHAPTER⟩ element did not include attributes other than ID, so a TYPE attribute could not be added for clarification. One of the dissertations had 150 illustrations and figures. There was no clear indication of what structure within the ETD DTD to use for the images, though the ⟨MM⟩ element (multimedia object) is a likely candidate. It was, however, not referenced in any content model other than its own, so that applications of ETD-ML could not use the ⟨MM⟩ element. In sum, the ETD DTD, as I found it in early March 1997, was insufficiently tested and had enough problems that it could not be used for this project.

This DTD underwent some revision in early 1997, and version 1.0 was released that March. This version fixed many problems with the beta by adding various floating elements to the ⟨ETD⟩ content model and including multimedia objects in other content models. I had completed the initial project by this point, but further electronic dissertation projects may be able to use this DTD successfully. I used the Text Encoding Initiative (TEI) DTD an alternative. I am most familiar with this DTD, plus it has the flexibility to deal with the complexities of many documents (including dissertations) and has been used by many universities and other organizations for text markup so a body of example documents, documentation, and a user community had developed. Markup of all four dissertations was done using the TEI Lite DTD, a subset of the full TEI DTD.

TEI Markup

The process used to mark up each dissertation varied depending on its length and complexity. I began with Tepper's dissertation, as it was the shortest, had no images, and had a reasonable number of footnotes. For this document, I used the SGML tools bundled with WordPerfect 7. It was a simple process of cutting and pasting from the Rich Text Format (RTF) text created from the MS Word for Macintosh version into the SGML document instance. The only difficulties lay in inserting lines of poetry, as WP's software did not have the split function that other SGML editors do. The split function allows the user to surround a larger portion of text as a particular element then split that section into smaller versions of the same element. In WP, each line had to be separately tagged with an ⟨L⟩, making for a tedious process.

Ruddy's dissertation was both longer and had more extensive footnotes than Tepper's. It also included Middle English characters, such as the thorn. Due to the substantial number and size of the footnotes, I saved the document as RTF, then used a Perl program to automatically mark up the text. This processing relied on distinguishing among the various RTF codes, which start with a curly brace then the codes for describing the text from that point to the ending curly brace (e.g., /footnote Source cited above.). Unlike SGML, the ending marker is generic, not indicating the element to which it refers. Because of this, some guesses had to be made on where the footnotes ended and some notes were inserted by hand.

Wheeler's dissertation on agricultural economics had more than 50 complex equations. The document was created in WordPerfect 7 so the formulas were done in WP's equation editor. To avoid rewriting the equations in SGML, I chose instead to use WordPerfect's automatic conversion to HTML. WordPerfect changed the equations to GIF images and the other formatting to HTML codes. These codes were regular and more easily identified than the RTF codes used in processing Ruddy's dissertation. Thus, with another Perl program, I was able to change the markup from the HTML DTD to the TEI Lite DTD. This required some cleanup and addition of elements, though hand-processing was quite limited in comparison to previous efforts. ISO characters were successfully changed from internal coding to character entity references.

Price-Wilkin’s dissertation presented the most complex challenges. It included several indices in the front matter and appendices in the back matter. Between these were an introduction, four chapters, and a conclusion, followed by many figures, illustrations, and tables. Each figure and illustration was provided in three forms: a thumbnail, 100 dpi images for on-screen viewing, and 300 dpi for printing. I chose to use the 100 dpi images in the SGML version to decrease the download time. TEI Lite has apparatus to enclose a thumbnail image in a reference to the larger image, but I chose not to do this to simplify the processing and re-conversion to HTML for non-SGML browsers. Again, I used WordPerfect as a first step to convert from RTF to HTML. As the HTML produced by WP is quite generic (e.g., what should be a ⟨H1⟩ in HTML is converted as ⟨P⟩⟨STRONG⟩), it was difficult to identify such structures as bibliographies, lists, and quotes. Due to the length of this dissertation and time constraints, some of these structures remain as ⟨P⟩s.

Why Use SGML?

SGML has many general advantages over plain or word-processed text. First, SGML files are usable on multiple platforms and no one software vendor controls the underlying language used. It lacks the proprietary coding that makes word-processed documents difficult to transfer between applications and platforms. The content of an SGML document is separated from its format, so the text can be rendered in different ways for different needs, platforms, and display methods (print, etc.). SGML is also often used as an archival format and for document reuse and repurposing. The use of SGML markup on dissertations allows far more complex searching than unstructured, wordprocessed text. For fully marked up documents, searches can be made on bibliographic citations (marked as a ⟨CITE⟩) or such citations could be extracted from each dissertation to create a citation database as a secondary product. Logical divisions within the text can be marked up and this structure utilized for retrieval of smaller portions of a document to reduce download time. As XML (a simple dialect of SGML for use on the Web) comes to the desktop, the raw SGML will become even more useful as it will translate to this new system equally well.

Markup Variation

A problem with using SGML stems from the variety in markup that can be produced by authors and that is allowed by the DTD. Footnotes in TEI exemplify this problem, allowing an author to tag notes following a pointer to that note, at the end of the chapter, and the end of the document. A way to avoid significant variation in applying markup is to have a central office for converting word-processed dissertations to SGML. At UM, there is already standardization required in preparation and formatting of dissertations, and printouts submitted to the UM graduate school are reviewed for compliance with these standards. With SGML, a stylesheet attached to the document would impose these formatting rules. The SGML DTD would impose some restrictions on how markup could be applied to a dissertation, and this markup would be reviewed.

Project Implementation

Since June 1997, the University of Michigan ETD project has progressed under two separate individuals. Between June 1997 and May 1998, Elizabeth Shaw, Digital Library Production Services (DLPS) special projects librarian, built on the initial findings of Janet Erickson, re-encoding the Ruddy and Price-Wilkin dissertations, and partially converting and encoding a new dissertation by David Meyer, a recent graduate from the UM Department of Economics. Shaw established more efficient mechanisms for converting and pre-tagging dissertations, and for analyzing the encoded documents, that together form the foundation of the current research process. Matt Stoeffler continued the project in June 1998. Stoeffler completed the Meyer dissertation, and has since encoded two additional dissertations from recent UM graduates in the departments of English Literature (Catherine Paul) and Mechanical Engineering (Jeannine Bos). A first draft of a working DTD has been completed, and it will be continuously revised as we test additional documents. DLPS also now has a relationship with the Graduate College to provide new dissertations at an increased rate, giving us enough data to sufficiently test the DTD over the next fiscal year.

Overall Goals and Methodology

The continuing goal of the UM ETD project is to develop and test a TEI-based DTD for ETDs across a variety of documents, with the express aim of arriving at a minimal subset of the TEI sufficient to effectively represent the structural complexity of dissertations, and yet simple enough that it can also provide a basis by which future documents could be authored in XML. It is hoped that by the early months of 1999, developments in new or existing word processing software will make it possible for a Ph.D. candidate to write a dissertation in SGML or XML, using the test DTD.

Currently, the test DTD exists as a work in progress. The element set consists of some 82 elements, all drawn from the TEI Lite DTD. Though there are some application issues that have emerged based on those dissertations already encoded, no new tags have been added to the original TEI element set. Our temptation is to resist addition of any elements until or unless document analysis confirms the need for new extensions of the TEI. In the meantime, we will convert and encode between 20 and 30 more documents through June 1999 using the following process: 1. the graduate college approaches a Ph.D. student at the time of final format review and obtains a finished draft on disk, which is sent to DLPS along with a print copy; 2. the document is converted to RTF format; RTFtoHTML and

Perl are used for pre-tagging; 3. an employee finishes the tagging, then another employee reviews the tagging, making annotations, and gives it back to the special projects librarian for final adjustments; 4. the finished document, along with the other finished dissertations, is run though a Perl script that counts the instances of all elements and their attribute values by document, and compiles an aggregate list of the elements across documents; 5. based on continued document analysis and results of the analysis script, the DTD is evaluated to see if there are poorly used or underutilized elements that should be removed from the final tag set. The final tag set may represent an additional 20 percent reduction in the number of elements, though no hard criteria for eliminating elements have yet been determined.

Converting from RTF to SGML

With but one exception, all of the recent documents that have come to us for encoding have been in Microsoft Word format on Mac or PC platforms. All are saved in RTF, which are converted to rough SGML via a two-step process. In the first step, DLPS currently uses a shareware filter application called RTFtoHTML (version 3.93) that converts the RTF document to an SGML-like HTML. Plain text library files that map paragraph and character styles in the source RTF to corresponding HTML tags make RTFtoHTML somewhat configurable. RTFtoHTML renders any embedded image into an ⟨IMG⟩-link (and on PC platform, converts the image to Windows Metafile (WMF) format), and renders all footnotes into anchor links to the note text. In the second step, several Perl scripts convert the ⟨IMG⟩ links to figure-entity relationships, align the footnote anchors into in-line ⟨NOTE⟩ elements, and filter the remaining HTML tags to corresponding SGML tags.

Issues and Problems in the Conversion Process

A. CONSISTENT APPLICATION OF PARAGRAPH STYLES

The primary weakness in the conversion process is that its effectiveness is largely dependent on the consistent application of Microsoft Word styles throughout the source document. With the Ruddy dissertation, the style sheet was applied quite extensively and consistently so that is was easy to identify main and subordinate structural divisions and assign them directly to DIV1 and DIV2 elements (first and second level divisions of a TEI document), as appropriate. The Meyer dissertation, however, did not apply the paragraph styles nearly so consistently, so that the filter was not able to differentiate content within the document very effectively. The variability with which authors apply paragraph styles in word processing documents presents a significant problem to any effort to institute a conversion program on a large scale, since it would necessarily push the labor costs of correction up too high for each document. Such a concern also underscores the importance of achieving a DTD that is not so complex that general authors would find it too difficult to apply effectively, and so partially defeat one main advantage of using SGML as a publishing format.

B. Inclusion of binary image files

Another conversion issue is the inclusion of binary image files. All image files, regardless of original format, once embedded in MS Word and converted via RTFtoHTML (in a PC environment) are converted into WMF format files with number-based filenames. Since WMF is not a universal format, we must first convert them into GIF or JPEG (or another appropriate) format.

A related conversion issue involved the Bos and Meyer dissertations, both from mathematics-intense disciplines. Each document included several mathematical formulas which were rendered as combinations of text or binary images created in MS Word's equation editor. As a conversion issue, the presence of such mixed-media formulas without identifying paragraph styles forced us to convert all embedded images as inline, and separate block images by hand, which takes considerably more time. The Bos dissertation also presented a pressing problem in the form of third party images imported into MS Word that did not make it cleanly through the conversion process. These images were created by the author in MATLAB (Mathworks, Inc.) from computational input and saved as encapsulated Postscript (EPS). The EPS was corrupted in MS Word so that it was effectively unusable. In the end, we had to request original files from the author; some were later scanned from the printed document. Such issues point to a significant advantage of "open" format authoring environments. Presumably, any future XML-capable word processor will facilitate, not complicate, the inclusion of external media into the current document instance according to the DTD in an appropriate format for storage and public distribution. Since there is strong likelihood that external or third-party media files will make their way into doctoral dissertations, the success of an XML-based dissertation authored in word processing applications may partially rely on the ease with which the application makes this possible.

Markup Issues

Conversion issues aside, there are a number of markup issues concerning how best to represent certain semantic or stylistic structures in the document instance that are still not firmly resolved. Among those issues, the most pressing concern the inclusion of formal equations or formulas and the multiple forms of bibliographic references, especially in the humanities fields.

A. Formal equations in mathematics-intensive documents

The Bos and Meyer dissertations, as mentioned above, presented a conversion problem when formulas were created as mixed-media, part text and part image. But such rendering of data in part-binary format raises an additional issue of how such content should be properly marked up in the first place to represent its semantic and stylistic nature. The Meyer and Bos dissertations required that the mixed-media be contained in a single element to reflect their semantic structure. Shaw's approach in the Meyer dissertation was to alter the TEI Lite DTD, changing the content model of the FORMULA element to include PCDATA, hi, figure, and label, and creating a new element, FORMFRAG, that would contain a portion of an entire formula that might be referred to later. Though this approach had the advantage of capturing the internal structural components of the content, it was felt that the distinctions between FORMULA and FORMFRAG seemed too complicated for the general author. The current approach taken in the Bos dissertation was to enclose such formulas within the Q element, with a TYPE attribute value of "formula," though the temptation is to migrate this to the SEG element with similar TYPE values. The SEG content model already has capacity for potential subelements without alteration. It seems an appropriate element to use for including structural components not otherwise reflected in the TEI Lite DTD, and so would be generalizable for marking up other objects unique to fields or individual disciplines. This issue is not fully resolved.

B. BIBLIOGRAPHIC CITATION

Another difficult markup problem concerns how multiple forms of bibliographic citations should be marked up. It is an issue of establishing style guidelines to help a general author mark them. Common practice in humanities scholarship, as reflected in the Paul and Tepper dissertations, is a case in point. Bibliographic references appear in humanities texts in multiple forms. Complicating the issue further is that many footnotes make multiple references intermixed with other annotations. Which if any of these forms should be recognized as a bibliographic citation marked up as a BIBL, and, if so encoded, to what degree should subelements (i.e., AUTHOR, TITLE, etc.) be encoded. In the Tepper dissertation, citations in all forms that were contained within footnotes were tagged as BIBLs, but no subelement tags were applied. For those instances where multiple citations occurred in one footnote, the distinction between what was a BIBL and what was not, or where one stopped and another began, seemed difficult to decide consistently. As a result, the approach taken in the Paul dissertation was not to contain any in-text citations within BIBL tags, and to let them stand as simple note text with highlighting supplied by HI elements. Part of the rationale here is that most, if not all, bibliographic citations will likely be repeated in a bibliography or other list of sources and that the proper structure of the full citation will be captured there, so that if someone wants to perform searches on the document for a particular string as "title" they will still achieve the desired result.

C. Other markup issues

There are some lesser markup problems that are not so clearly defined yet. Among them is the question of how to encode bibliographic citations to networked resources where the part or all of the citation is in fact a URL. This issue surfaced in the Tepper dissertation, and our tentative response has been to encode the URL within REF elements, with TYPE attribute value of "URL." We will revisit the issue when the project is closer to an online XML application. Another issue that recently emerged is that of intellectual property rights. In this case, the Paul included some 60 images as illustrations in her dissertation, the majority of which she had permission to use but not distribute. Currently we have no capacity for assigning attribute values corresponding to availability of subelements within a document for use in filtering those "private" elements out of the document as it is delivered to the user. Our current workaround has been to create two versions of the same document, one substituting descriptions in place of those figures that are not for public distribution.

Conclusion

In both initial and subsequent investigations of electronic dissertations at UM, we have found that the TEI Lite DTD can capture the essence of a doctoral dissertation. Many problems remain to be solved, especially regarding complex mathematical formulas, dealing with inconsistent style usage, varied bibliographic citation, and intellectual property. We hope that the work described here will advance the cause of electronic submission of dissertations and theses using archival, nonproprietary, and format-neutral SGML markup.

References and Resources

Bos, Jeannine. *A Study of Heat Transfer and Flow with Phase Change in Deep Penetration Welding.* Dissertation (Ph.D.) – University of Michigan, 1998.

Erickson, Janet. *Mock-up of The University of Michigan Dissertation and Thesis Library*. http://dns.hti.umich.edu/misc/diss.example/index.html

Erickson, Janet. "An SGML/HTML Electronic Thesis and Dissertation Library". *DRAFT*. http://www-personal.umich.edu/˜janete/tei_etd.htm

Fox, Edward A. et al. "National Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources". *D-Lib Magazine*, Sept., 1996. http://www.dlib.org/dlib/september96/theses/09fox.html

Fox, Edward A. et al. "Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources". *D-Lib Magazine*, Sept., 1997. http://www.dlib.org/dlib/september97/theses/09fox.html

Kipp, Neil A. *User's Guide to Electronic Theses and Dissertation Markup Language (ETD-ML)*. http://etd.vt.edu/etd-ml/userguid.htm

Kipp, Neil A. *Document Type Definition for Electronic Theses and Dissertations*. http://etd.vt.edu/etd-ml/dtdetds.htm

Kirschenbaum, Matthew G. *Electronic Theses and Dissertations in the Humanities: A Directory of Online Reference and Resources*. http://etext.lib.virginia.edu/ETD/ETD.html

McMillan, Gail. "Electronic Theses and Dissertations: Merging Perspectives at Virginia Tech". *Cataloging & Classification Quarterly*, 22(3–4) (1996), 105–25. A 1995 draft of this article is available at http://scholar.lib.vt.edu/theses/GailsCCQarticle.html

Meyer, David. *Essays on Quality and Product Differentiation*. Dissertation (Ph.D.) – University of Michigan, 1996.

Networked Digital Library of Theses and Dissertations. http://www.ndltd.org/

Paul, Catherine Elizabeth. *Poetry in the Museums of Modernism: W.B. Yeats, Ezra Pound, Marianne Moore*. Dissertation (Ph.D.) – University of Michigan, 1998.

Price-Wilkin, Rebecca Mary. *The Late Gothic abbey Church of Saint-Riquier: An Investigation of Historical Consciousness*. Dissertation (Ph.D.) – University of Michigan, 2 volumes, 1997.

RTFtoHTML for converting Rich Text Format documents to an SGML-like HTML. Available at http://www.sunpack.com/RTF/

Ruddy, David Wilmot. *Scribes, Printers, and Vernacular Authority: A Study in the Late-Medieval and Early-Modern Reception of Mandeville's Travels*. Dissertation (Ph.D.) – University of Michigan, 1995.

Tepper, Michele Eden. *The Mind of His Own Country: Embodiments of National Culture in Modernist Literature (T.S. Eliot, Gertrude Stein, William Carlos Williams)*. Dissertation (Ph.D.) – University of Michigan, 1998.

UMI. *Submitting Your Dissertation or Master's Theses in Electronic Format*. http://www.umi.com/hp/Support/DExplorer/prepare/submit.htm

University of Michigan TEI ETD Document Type Definition. http://dns.hti.umich.edu/dissert/dtd/teietd.dtd

University of Waterloo Electronic Thesis Project Survey Results http://library.uwaterloo.ca/˜uw-etpt/

Virginia Tech Electronic Thesis and Dissertation Project. http://scholar.lib.vt.edu/theses/ (Scholarly Communications Project C Virginia Tech Electronic Theses and Dissertations); http://etd.vt.edu/submit/ (submission information); http://www.theses.org/ *or* http://www.dissertations.org/ (ETD Digital Library); http://www.ndltd.org/listserv/ (ETD listservs).

Wheeler, William James. *Discounting and the Evaluation of Global Warming Policies (Hyperbolic Functional Form, Rate of Time Preference)*. Dissertation (Ph.D.) – Pennsylvania State University, 1997.