# Automation Matters

*Paul Hedges*
*State Historical Society of Wisconsin*

## Standard Generalized Markup Language (SGML): Part Two

In the previous column I attempted to define Standard Generalized Markup Language. I hope that I accomplished my goal. In this column I would like to pick up where I left off and talk about a couple of SGML approaches that may be of interest to archivists. Before going any further though, let me briefly recap what SGML is. SGML is an important electronic document technology that can help in the creation of electronic documents in a standard and accessible form despite hardware and software changes over time. SGML is a technology that allows for the creation of a document type definition (DTD) that defines the structure and content of an electronic document. For example, a memo and letter each have a specific structure and content that we can identify and formalize. Sections in either of these documents might include a recipient's address, a body of text, date of document, or salutation. Other documents, such as telephone books, airplane maintenance manuals, dictionaries, and archival finding aids, have a much more complex structure requiring complex DTDs.

Once the structure and content are defined, unique codes are created to mark the corresponding portions of the text. These codes are called "markup codes" or "tags." A markup code is usually represented within "<" and ">" characters and paired as begin codes and end codes. For example, the tags <title> and </title> would mark the text between them as the title (or whatever the DTD defined it to be). Other examples include <chap>...</chap>, <verse>...</verse>, <author>...</author>, and <repository>...</repository>. Each example uses the "/" in the second tag to specify that it is the end tag. A final, important aspect of SGML is that the documents are plain text files (ASCII or ANSI). This allows the electronic document to be stored, transferred and analyzed by all current (and future) computer hardware and software applications.

In this column I have decided to focus on two research projects that use SGML as a tool to develop standards for electronic text: The Berkeley Finding Aid Project and the Text Encoding Initiative. The Berkeley Project is unique in that it is attempting to define a standard used every day in archival institutions, the finding aid. The Text Encoding Initiative is unique in that it is defining standards for electronic texts that may, one day, be a part of archives' collections.

### Berkeley Finding Aid Project

According to the project description, "[the] Berkeley Finding Aid Project is a collaborative effort to test the feasibility and desirability of developing an encoding standard for archive, museum, and library finding aids." To date, the project has defined a prototype finding aid, identified the various structural and content components, and developed a draft DTD. Staff of the project are now tagging archival finding aids that they have received from various archives and museums. The next step is to test these sample finding aids on a computer that reads SGML codes. One component of the testing will be to make the SGML encoded finding aids available to other institutions via the Internet.

SGML formatted finding aids can provide several benefits. First, finding aids created using SGML can be accessed, searched and manipulated on any computer hardware and software now or in the future. This goes a long way toward assuring archivists that their effort in converting paper finding aids into electronic form is a good long term investment. Second, SGML can create DTDs for complex, hierarchical documents. This is a crucial requirement for finding aids because archivists use many levels of headings, subheadings, and sub-sub headings in their container lists. The draft Berkeley Project finding aid DTD has been created to accommodate multiple levels of headings. Third, the document type definition in SGML does not define the appearance of a structured document, only the internal structure and content. What this means to you, the user, is that one data file containing a finding aid encoded according to the Berkeley SGML DTD could be used to produce several different versions of the finding aid—a reading room version, an abstract version, multi-document index, and to create either plain text files for your Gopher Site or instant HTML formatted documents for your World Wide Web site. Finally, encoded finding aids can ultimately be linked to digitized versions of primary material, allowing a researcher to navigate from a bibliographic record to the finding aid and then to the primary material itself.

### Text Encoding Initiative (TEI)

The Text Encoding Initiative began in late 1987 at a planning conference sponsored by the Association for Computers and the Humanities (ACH). At this conference, representatives of text archives, scholarly societies, and research projects met to discuss the feasibility of developing a standard encoding scheme for electronic humanities text. The outcome of this meeting was a set of principles that determined the course of development for the Initiative. In low tech speak, TEI was to identify an encoding method, develop basic DTDs for a variety of humanities texts, and then provide guidelines for using the new DTDs. To date, TEI has developed definitions and standard markup tags for the following document types: prose, verse, drama, tran-



Generated at Iowa State University on 2021-02-11 17:57 GMT / https://hdl.handle.net/2027/mdp.39015068967846
Creative Commons Attribution-NonCommercial / http://www.hathitrust.org/access_use#cc-by-nc-4.0
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

scription of spoken texts, dictionaries, and terminological data files. Supplemental markup tags have also been defined to permit the designation of hypertext links, names and dates, graphics, analytics, and specific tags for manuscripts, analytic bibliography, and transcription of primary sources. A final, and very important goal of the initiative is to add to the guidelines and DTDs over time while maintaining downward compatibility. This will assure that all text encoded today will be readable by newer TEI standards in the future.

To date, the Oxford University Oxford Text Archive, the University of Virginia Electronic Text Center, the University of Michigan Humanities Text Initiative, and the Brown University Scholarly Technology Group all use the TEI Guidelines and have placed a significant amount of scholarly text into TEI compliant SGML format. The effort has focused on taking ancient texts, reference material and modern literary works, encoding them, and placing them in database programs. The result is a corpus of documents that can be searched, analyzed for content and structure, and formatted at will. Readers with Internet Gopher or World Wide Web access make sure to check out each university for examples of their SGML encoded works.

### What's Next?

There is much potential here, but there is also much hard work ahead for most Archives. The software tools for creating and editing SGML documents are still expensive and difficult to use. This will change as SGML becomes more mainstream and competition increases. For example, WordPerfect recently announced that they will soon release a SGML-aware version of their WordPerfect for Windows 6.1 product, and other low cost products are slated for release in the next six months.

Although the tools will be available, an Archives must realistically look at the benefits of standard automated finding aids versus the costs of training, conversion, and long term maintenance. As one who is comfortable with the technology I must admit I am excited by its possibilities, but only time will tell if SGML encoded documents will have a lasting role in archival collection management.

## Moody Files Opened

*(continued from page 1)*

Johnson, and many others.

The records were given to the Archives by Moody Church in 1986. They have now been processed by the Archives staff and, as of December 1, 1994, are officially open to the public. The collection includes 65 boxes of files, as well as hundreds of photographs and glass slides. The earliest documents date from 1872, just a few years after the church's founding in 1864, and the most recent ones are from the 1980s, although most of the material in the collection is from a period roughly 1915–1950.

The church started as an informal Sunday School organized by Dwight L. Moody in the late 1850s. He and others incorporated it as a church in 1864. The building on Illinois Street did not survive the great fire of 1871, but the church continued to thrive in a variety of homes until settling in 1925 into its present structure at the corner of LaSalle and North Avenues. Many prominent Christian leaders spoke from its pulpit, and the congregation ran Gospel missions and soup kitchens, in addition to a variety of outreach programs that it developed for social and ethnic groups in the city.

People desiring more information about the collection should contact the Archives at (708) 752-5910 or through e-mail at bgcarc@david.wheaton.edu.

The Archives of the Billy Graham Center collects material on nondenominational North American Protestant efforts to spread the Christian Gospel. Its hundreds of collections include the records of mission agencies, evangelistic organizations, and service associations; the private papers of individual Christian workers; oral history interviews; and the files of conferences and congresses.

## Indiana University Receives Grant

*(continued from page 1)*

procedure statements for appraising, describing, accessing and preserving IU's electronic records.

One of the primary objectives of the project is to develop specific methodologies and practices which will add to the knowledge pool regarding the proper management of electronic records. Phil Bantin emphasized this goal when he noted that "managing electronic records is a number one priority for the archives profession.

One primary strategy for advancing the profession's understanding of this topic is to develop usable models and generalized results. The IU project places strong emphasis on developing model policies and procedures that can be applied beyond the IU community."

Another primary feature of this project is its multi-disciplinary nature. The project has been designed to involve IU personnel from a variety of disciplines in active and advisory roles. Gerry Bernbom, Assistant Director in the Office of Information Technologies, noted: "In the research phase, the project will provide an opportunity to make a contribution to Archives and Data Administration through the synergy of knowledge and practice of these two professions."

The project is jointly sponsored by the University Archives, University Computing Services and the University Libraries, and supported by Indiana University's information policy coordinating committees: the Committee of Data Stewards, and the Committee on Institutional Data. In addition, faculty, librarians and administrators from the University Administration, University Computing Services, the School of Library and Information Science, the School of Public and Environmental Affairs and the University Libraries' Automation, Preservation and Cataloging Departments will serve in advisory roles.



Generated at Iowa State University on 2021-02-11 17:57 GMT / https://hdl.handle.net/2027/mdp.39015068967846
Creative Commons Attribution-NonCommercial / http://www.hathitrust.org/access_use#cc-by-nc-4.0
Digitized by Google
Original from UNIVERSITY OF MICHIGAN