

2018

On The Frontlines of Richer Metadata: Technical Services and TEI

Elizabeth Hertenstein

Bowling Green State University, eherten@bgsu.edu

Julie Rabine

Bowling Green State University, jrabine@bgsu.edu

Mathew Sweet

Bowling Green State University, sweetm@bgsu.edu

Follow this and additional works at: https://scholarworks.bgsu.edu/ul_pub



Part of the [Library and Information Science Commons](#)

Repository Citation

Hertenstein, Elizabeth; Rabine, Julie; and Sweet, Mathew, "On The Frontlines of Richer Metadata: Technical Services and TEI" (2018). *University Libraries Faculty Publications*. 54.
https://scholarworks.bgsu.edu/ul_pub/54

This Article is brought to you for free and open access by the University Libraries at ScholarWorks@BGSU. It has been accepted for inclusion in University Libraries Faculty Publications by an authorized administrator of ScholarWorks@BGSU.

On The Frontlines of Richer Metadata: Technical Services and TEI

Elizabeth M Hertenstein

Bowling Green State University

Jerome Library, Rm 233

1001 E Wooster St

Bowling Green, OH 43403

eherten@bgsu.edu

Julie L Rabine (Corresponding author)

Bowling Green State University

Jerome Library, Rm 234

1001 E Wooster St

Bowling Green, OH 43403

jrabine@bgsu.edu

Mathew C Sweet

Bowling Green State University

Jerome Library, Rm 231

1001 E Wooster St

Bowling Green, OH 43403

msweet@bgsu.edu

Introduction

For many libraries, traditional technical services work has been steadily shrinking along with the size of print collections, as the automation of processes continues to increase. Much of our focus as library technical staff is now devoted to electronic resources, but attention must also be paid to innovative areas of cataloging and metadata such as linked data and the digital humanities. The digital humanities, being the practice of using technology to leverage humanities research, has been used locally in a variety of forms including collaborations on digital exhibits with students and faculty, and projects on text analysis.

The authors have identified interest in the digital humanities as a growth area for the library's technical services department and found the two to be naturally compatible. A large number of tasks related to the digital humanities utilize the technical services skill set, including metadata creation, experience in manipulating data using a variety of encoding formats, and work with records where attention to detail is critical. The digital humanities can bring us new ways to fulfill our traditional mission to make information accessible, and even take it a step further by collaborating to create and display new knowledge. Work with the digital humanities and linked data can also create opportunities for external outreach, which is an increasing aspect of technical services work at our institution.

These areas present legitimate ways to provide meaningful work for technical services staff, alongside common tasks such as copy-cataloging, batch loading records, and troubleshooting access issues with electronic journals and databases. As the authors will demonstrate, technical services departments can participate in digital humanities and linked data projects to bring unique local materials into the public eye, even with a very small staff.

At Bowling Green State University Libraries, the Collections and Technical Services staff have been partnering with the Center for Archival Collections to create Text Encoding Initiative (TEI) documents from a selection of archival materials, starting with a collection of

letters from World War I. This article will look at the challenges of constructing workflows for a project which comes under the purview of both the archives and the technical services departments, with particular focus on the questions and discussions that informed crucial decisions directing the project.

Literature Review

Academic libraries have been part of the dialogue surrounding the development of TEI since its inception, and they continue to be closely involved in its evolution (Cantara pt. 2, 2005). Librarians were among the scholars and specialists gathered at Vassar College in 1987 to create the encoding guidelines which would become the Text Encoding Initiative (Cantara pt. 1, 2005), and since 2000 the interests of libraries and librarians have become a greater focus for the principal directors behind TEI (Barney, 2012). Cataloging and metadata handbooks have covered TEI since at least 2003 (Caplan) and continue to do so to the present (Eden, 2004; Smiraglia, 2005; Liu, 2007; and Zeng, 2016). Today the TEI-C Libraries Special Interest Group continues to update and maintain a collection of best practices for TEI in libraries (Best Practices 3.0, 2011). As of this writing, a task force has prepared a new revision that is now up for comment (Best Practices, 2017).

The Digital Humanities (DH) was listed as one of C&RL News' top trends in academic libraries for 2014/15. Writing in the summer of 2014, the ACRL Research Planning and Review Committee saw DH as a key area for academic libraries to collaborate with the larger campus community (ACRL, 2014, p. 299).

Little has been written specifically concerning library technical services departments coding with TEI. In fact, at least among ARL libraries, the trend has been for "[n]on-MARC metadata [to be] created by non-cataloging librarians and paraprofessionals from other departments as well as those who work in technical services," according to a survey done by Fleming, Mering, and Wolfe (2008, p. 3). The most common departments to be involved with

metadata creation outside technical services were archives and special collections, digital initiatives, and systems (Fleming et al., p. 7). A more recent survey of ARL libraries conducted in 2011 showed that "with the exception of a few well-known programs, most library-based DH is being done in a very piecemeal fashion." As a result, the "success of library DH efforts often depends on the energy, creativity, and goodwill of a few overextended library professionals and the services they can cobble together" (Posner, 2013 p. 44). Posner goes on to discuss the importance of administrative support needed to lift DH programs above the level of personal projects and into the realm of library work. One particularly notable example in the literature of a technical services department working with is TEI: Case Western Reserve's project to encode their collection of historical texts related to the history of Cleveland, Ohio and the surrounding area (Wisneski & Dressler, 2009). Following this article, very little subsequent research has been published on the topic.

Background

Bowling Green State University (BGSU) is a mid-sized university located in Bowling Green, Ohio and enrolling approximately 15,000 FTE undergraduate, masters, and doctoral students. Jerome Library is the main campus library, housing over six million items and garnering national recognition for its extensive popular culture and music collections. In addition, the Center for Archival Collections (CAC) boasts several unique collections, including the Historical Collection of the Great Lakes, The National Student Affairs Archive, the rare books collection, University Archives, and manuscript collections related to the history of Northwest Ohio.

The Collections and Technical Services (CATS) department covers acquisitions, cataloging, collection development, electronic resources, serials, and processing of government documents. The unit consists of eleven staff members, including four library

faculty, one administrative staff, and six classified staff. Jerome Library has a library IT unit (LITS) consisting of three administrative staff and one classified staff.

History of the project

The idea of TEI was first introduced to University Libraries by the Metadata Librarian, who learned about TEI application, standards, and procedures while attending Rare Books School at the University of Virginia, Charlottesville. The Emerson Family Papers were the UL's first candidate for TEI encoding. The collection, acquired by the CAC from Charles Apfelbaum in 2005, consists of twenty letters written by two U.S. Army privates, brothers Earl and Scott Emerson. The letters were sent from various placements around France and Germany in 1918 and 1919. The collection was chosen because of its cross-discipline appeal to historians, linguists, and scholars; the relatively small size of the collection; and its lack of accessibility at the time.

Text Encoding Initiative

TEI is a way to encode text which allows a wide range of metadata to be embedded. The guidelines maintained by the Text Encoding Initiative Consortium (2016) present a wide range of possibilities for text markup, with different aspects of the system adding different functionality, as might suit the project in hand.

TEI involves the use of tags to add information about specific words, phrases, and sections. For example, the following sentence uses tags to identify that the word “states” is here used to refer to a country:

```
Some of the volunteers from<lb/>
the <country>states</country> just landed<lb/>
here and they will<lb/>
```

The tag <lb> shown in this example denotes a line break, accurately representing the original layout of the letter which is being digitally represented. The TEI in this example can

be developed further, so that the country which is identified by the word “states” can be identified more specifically as the United States:

```
Some of the volunteers from<lb/>
the <country key="United States">states</country> just landed<lb/>
here and they will<lb/>
```

Embedding the information in this way allows people searching the letters the option to identify any or all uses of the word “states” which are used in the letters to denote the country the United States of America. Furthermore, all references to the United States can easily be identified, regardless of the original author’s presentation or spelling. TEI elements can be nested, so that several pieces of information pertaining to a single instance of a particular word or phrase can be expressed in a line. For example, if the word “states” were misspelled in the sentence above, it could be expressed using TEI in this way:

```
Some of the volunteers from<lb/>
the <country key="United
States"><choice><orig>staites</orig><reg type="misspelling">states</reg></choice></coun
try> just landed<lb/>
here and they will<lb/>
```

The <choice> element here stipulates that the original text in the document used the spelling “staites,” whereas the regular spelling of that word would be “states.” In this way, the original appearance of the letter is maintained, but with a potentially useful correction, as well as an embedded definition. The way that these aspects are displayed on a website will be determined by the way the stylesheet on the hosting website is coded to process the elements in the XML file.

The TEI guidelines list hundreds of elements and ways in which they may be used to mark up a document. It is important to note that TEI is not an objective process; there may be several ways of integrating the same information, and there may be varying degrees of detail applied. Using the example above, another person applying TEI markup to the same text may arrive at the following:

```
Some of the volunteers from<lb/>
the <location><placeName xml:id="United States"><sic>staites</sic>
<corr>States</corr></placeName></location> just landed<lb/>
<location><placeName xml:id="France"> here</placeName></location> and they will<lb/>
```

Structured this way, the information embedded is essentially the same, but the method is clearly very different. With this in mind, it can be prudent to specify a limited schema for a project in order to increase consistency across all encoded documents, and thereby increase consistency in search results and metadata presentation. A list of the tags ultimately chosen and used by the authors is further discussed below.

Process

The Metadata Librarian and Cataloging Coordinator planned to work together on an initial mark-up of a selected portion of the collection, with the assistance of students. After some experimenting, they agreed that given the normal margin for human error, transcriptions would benefit from review by persons other than the initial transcriber in order to maintain quality. To coordinate this process, the authors created an Excel spreadsheet with a row for each page of each letter in the collection. Columns indicated who transcribed the page, who reviewed the transcription, and who applied TEI mark-up. Other columns denoted specific details about the pages including the original author and recipient, the date it was written, and a further column for additional notes. This spreadsheet was especially important given the number of students who were rotated in and out of the project.

Although adding students to the venture helped move the work forward more quickly, problems arose with certain aspects of the process. It emerged that many of today's college students did not learn cursive writing while in elementary school, and this unfamiliarity meant that they would have difficulty reading parts of the letters which staff members were able to decipher with relative ease. Likewise, student workers tended to mark words as incorrectly spelled when they were unable to accurately distinguish the handwritten letters. Students also made transcription errors or introduced their own spelling mistakes into the documents, making it hard to determine whether certain anomalies were in the original letters or not. These issues were resolved through quality control by the professional staff supervising the project.

Lack of time created further problems. Students transcribed pages faster than professional staff could perform quality control. The resulting bottleneck in the workflow meant that feedback to students on their work was often delayed, leading to repeated errors on subsequent papers. At this point it was clear, that given the level of detail needed, more professional staff needed to be involved in the project to supervise and train students. A Library Associate was then invited to join the project. The addition of the Library Associate meant not only accurate transcriptions, but a third reviewer as well. To maintain a steadier workflow, the authors agreed to forego the additional speed provided by student workers.

Tagging and Encoding

The TEI guidelines (2015) provide comprehensive descriptions and instructions for use of TEI, including information regarding the 5 different encoding levels. The authors decided that level 4, "basic content analysis," best described the coding level appropriate for the World War One letters project. Level 4 analysis "...identifies hierarchy and typography, specifies function of textual and structural elements, and describes the nature of the content and not merely its appearance." (Best Practices 3.0, section 1, 2011.)

In order to determine which TEI tags would best suit the project, the Metadata Librarian and the Coordinator read through the letters, identifying the main concepts and themes, and considering which aspects might best be highlighted. With this information they were able to select the TEI tags most likely to be relevant and informative for these texts, and develop a limited pool of tags. The purpose of the pool was merely to promote consistency across different people's work, not to apply unnecessary limitations. With the pool completed, the Metadata Librarian and the Coordinator were ready to start coding letters. The complete pool of tags used by the authors may be seen in Appendix one. Throughout this process, a detailed policies and procedures document was created and updated, with the aim of providing an instructional source for training.

The transcriptions, initially created as text files, were imported into the XML editor Oxygen in order for TEI mark-up to be applied. With the time-consuming transcription work already completed, and the pool of TEI tags predetermined, the process of applying TEI mark-up was reasonably straightforward for most letters, and could be finished in a short period of time. After multiple problems with students as “coders,” staff agreed that it would be wiser to use CATS staff exclusively for this part of the project. Although each page would take longer to process, the error rate would prove to be significantly lower, and thus the pace of the review and correction phase was much improved.

The Metadata Librarian and the Coordinator initially attempted to create the required headers for the XML files, but it quickly became apparent that the technical expertise necessary for this coding was lacking, so the library IT unit (LITS) was consulted. LITS staff were able to script initial versions of these components, which could then be pasted into each document and edited as appropriate. A sample header, created by John Kloor, can be viewed in Appendix two.

Access and display

Once the XML files had been completed and reviewed, the authors met to discuss the options for presenting the documents publicly online. Omeka was initially proposed as a platform, its advantages being that it is open source (and thus free to use), and was already in use by the library for various online exhibits and displays. However, the Omeka TEI plug-in was found to be cumbersome and IT staff were not able to make it function satisfactorily. The authors also considered the possibility of using ContentDM, but the “lite” version available to OCLC subscribers was too limited for the purpose, and staff in LITS were unable to spare the necessary time to offer support for such an unfamiliar system; the full version was simply too expensive and shared the IT support issues. As a result of these discussions between the authors and LITS, it was concluded that a purpose-built website would be the best solution.

Given the in-house system, questions remained about page layout, design, and what sort of navigational options would be desirable. The authors agreed that a simple, clean aesthetic would best support the intention of highlighting the documents and the underlying TEI as well as allowing a variety of potentially disparate projects to be housed on the same site without incurring any stylistic clashes. The Library Associate put together a site design using Photoshop to make non-functioning images of how the pages might look, in order to visualize ideas and share them with the group. After a few adjustments, these were passed to the LITS department along with the individual image files utilized, in order to be compiled as a series of functioning webpages.

Various aspects of the project required close liaison with the LITS unit, and their input was invaluable to realizing the project. LITS staff set up the site and also created the complex stylesheets necessary to visually differentiate various aspects of the TEI encoding (Appendix two). In most cases, the words or phrases with TEI tags attached show in the transcription as highlighted in various colors according to the type of tag applied. Mousing-over these highlighted words and phrases reveals additional information besides the script.

Upon mouse-over, some elements are shown to be clickable by virtue of changing the cursor to a finger point. Clicking on these elements leads to additional information on a separate page. At the time of writing, this last type of coding is used exclusively for names which occur in the letters, and clicking on them leads to the relevant entry in the personography, but this tool has much potential, and will likely be used in other ways in later projects. The stylesheet, created by John Kloor, is too large to feature as an appendix but is available upon request to the authors.

The creation of the so-called personography, or taxonomy of all personal names referenced in the text, was an interesting project in and of itself. The Ancestry.com database is a rich source of biographical data, and searching it provided information including full names, birth and death dates, residences, and family relationships. Using this data, the authors were able to identify many individuals mentioned in the letters, such as aunts, uncles, and neighbors who were referred to only by a nickname. Searching the files of military records available also gave insight on the activities of various individuals during the war, such as military deferments and dates of service. The Emerson personography may be viewed in Appendix three.

TEI requires a robust search engine in order to take full advantage of the specific content tags, as opposed to the broader functionality of a general keyword search. At the time of writing, some fundamental questions remain about how the BGSU TEI project search page might operate. Ideally, users will be able to search for all instances of a particular name, place, or item, regardless of how they might be presented in the text. Users should have the option to search within specific collections, or all across all collections, and they should have the ability to see all examples of misspellings or modernizations, or all the ways in which the time of day is represented, and so on. All these considerations and more fuel the current

discussion about the search page. Once practical conclusions have been agreed upon, staff will consider design options for incorporating this functionality.

Conclusion

Text mining and other methods of digital humanities research have become increasingly popular among faculty and graduate students in humanities and social sciences fields at BGSU, and UL librarians have participated in this trend in various ways. For example, in one of the first DH projects in the library, librarians in the Browne Popular Culture Library have been working with an instructor to curate digital exhibits of primary source material found in the collection, as a class project. It seems reasonable to suppose that CATS librarians can also find an appropriate place in the digital humanities.

The Cataloging Coordinator, now also the Chair of CATS, has two chief reasons for bringing TEI encoding into the Technical Services department. The first is that as a university library, creating and sharing knowledge is a foundational part of our mission. Digitizing our materials using this form of metadata not only makes the individual documents more accessible, it enables their inclusion in a larger dataset for researchers to study in a new way. Our sets of WWI letters, for example, can be combined with others to form a critical mass of data for text mining to study history, genealogy, language, and even spelling changes. The digital humanities is essentially “big data” for humanities scholars, and TEI forms the basis for this type of scholarship.

The second reason for doing TEI in the library is to bring our technical services department into the 21st century. Much of our work is outsourced these days, and even more is automated. We get fewer and fewer print and tangible items every year. We automated much of our workflow with the idea that it would free our staff from the routine work and allow them to spend their time on more skilled activities. Projects like this one take great advantage of skilled staff, while providing interesting and engaging tasks. Such projects are

indicative of the type of work a library's technical staff might be best suited for in the near future, where application of research, acumen, and problem solving are core. Participating in the digital humanities is an important way for our technical services department to move forward, and what started as a research project has become part of our regular work.

Today, CATS librarians are participating in a number of different collaborations and ventures in addition to the TEI work described in the article. Librarians in CATS and the Music Library are collaborating to create a linked data project connecting the music and musicians who are part of the annual New Music Festival at BGSU. In fact, librarians are exploring how MEI (Music Encoding Initiative, a sister to TEI) could be used in conjunction with TEI on future projects. More cross-departmental projects have also been proposed, including working with the Browne Popular Culture Library on a TEI enterprise focused on the department's collection of pulp science fiction novels, a project that is still taking shape.

The authors' long-standing goal of reaching out to faculty outside the library is starting to bear fruit as well. In spring 2017 the Metadata Librarian consulted with two professors of Russian who were considering having their students encode a chapter of Anna Karenina with TEI as a class project. The authors more recently consulted with a Communications and Cultural Studies professor who is interested in using TEI to enhance the oral histories her students are documenting. This last project is of particular interest in that it will be the first time CATS staff have worked on encoding materials for a non-library-based initiative.

All in all, the initial experience of using TEI mark-up at BGSU has been extremely interesting and informative. By bringing together the CAC and CATS departments, a selection of documents that might otherwise have been largely ignored has been highlighted and presented in a way that maximizes their potential for interested parties. Use of TEI mark-

up has enriched the content of the documents, and made it possible for the documents to be searched in a number of unique ways.

As information professionals, we want to add enriched metadata that goes beyond Dublin Core; as faculty, we want to engage in scholarship and the creation of new knowledge; as librarians we want to engage with researchers, instructors, and students in new ways. TEI has provided the platform for us accomplish these goals.

The BGSU TEI web site may be viewed at <http://lib.bgsu.edu/tei>. The authors would like to give special thanks to John Kloor and LITS for all the work they have put into this project despite a packed schedule. As of this writing, the site is awaiting updates and the addition of new material that have been sent to LITS staff for processing.

Bibliography

- Barbera, M. (2013). Linked (open) data at web scale: research, social and engineering challenges in the digital humanities. *JLIS.it*, 4(1), 92-101.
- Barney, B. (2012). Digital Editing with the TEI Yesterday, Today, and Tomorrow. *Textual Culture*, 7(1), 29-41.
- Best Practices for TEI in Libraries: A Guide for mass digitization, automated workflows, and promotion of interoperability with XML using the TEI.* (2017) Retrieved October 31, 2017, from: <http://www.tei-c.org/SIG/Libraries/teiinlibraries/>
- Best Practices for TEI in Libraries (version 3.0, published October 2011)* (2011). Retrieved January 17, 2018, from: <http://www.tei-c.org/SIG/Libraries/teiinlibraries/main-driver>
- Cantara, L. (2005). Digital Libraries in the Humanities The text-encoding initiative: Part 1. *OCLC System & Services: International Digital Library Perspectives*, 21(1), 36-39.
- Caplan, P. (2003). *Metadata Fundamental for All Librarians*. Chicago: American Library Association.
- Eden, B.L. (Ed.). (2004). *Innovative Redesign and Reorganization of Library Technical Services: Paths for the Future and Case Studies*. Westport: Libraries Unlimited.
- Fleming, A., Mering, M., & Wolfe, J.A. (2008). Library Personnel's Role in the Creation of Metadata: A Survey of Academic Libraries. *Technical Services Quarterly* 25(4), 1-15.
- Liu, J. (2007). *Metadata and Its Applications in the Digital Library: Approaches and Practices*. Westport: Libraries Unlimited.
- Posner, M. (2013). No Half Measures: Overcoming Common Challenges to Doing Digital Humanities in the Library. *Journal of Library Administration*, 53(1), 43-52.
- Smiraglia, R.P. (Ed.). (2005). *Metadata: A Cataloger's Primer*. Binghamton: Haworth Information Press.

TEI: Text Encoding Initiative (2016). Retrieved November 15, 2017, from: <http://www.tei-c.org/index.xml>

TEI: Guidelines (2015) in *TEI: Text Encoding Initiative*. Retrieved November 15, 2017, from: <http://www.tei-c.org/Guidelines/>

ACRL Research Planning and Review Committee. (2014). Top trends in academic libraries: a review of the trends and issues affecting academic libraries in higher education. *College & Research Libraries News*, 75(6), 294-302.

Wisneski, R., & Dressler, V. (2009). Implementing TEI Projects and Accompanying Metadata for Small Libraries: Rationale and Best Practices. *Journal of Library Metadata*, 9, 264-288.

Zeng, M.L., & Qin, J. (2014). *Metadata*. 2nd ,rev. Ed. Chicago: American Library Association.