DATA PREPARATION FOR A DIGITAL HUMANITIES DOCUMENT: CHALLENGES AND OBSERVATIONS

Katia P. Mayfield **Emmanuel Grant** **Crystal Alberts**
Department of Computer Science **Department of English**
University of North Dakota
Grand Forks, ND 58201

Abstract

Digital Humanities is an interdisciplinary field that involves teaching, study and research in terms of computers and the different areas of Humanities. The purpose of this interdisciplinary field is to concentrate on the area of "born-digital" data, digitization of standard data, and their analysis. Organizations have been created to categorize materials that are considered to be born-digital, and to work with archival documents. These entities are expected to follow the Text Encoding Initiative standards. Several studies have been published in regards to Digital Humanities with mapping concepts from printed to digital information, data gathering, difficulties in producing metadata, and characteristics of documents archived in Digital Libraries. This paper focuses on describing a real life, step by step, process of converting a piece of media to a Digital Humanities accessible entity.

1 INTRODUCTION

Digital Humanities is an interdisciplinary field combining computational sciences and the different areas categorized as humanities, such as literature, anthropology, and history. Digital Humanities is also sometimes referred to as humanities computing. The purpose of this interdisciplinary field is to concentrate on the area of digitization and analysis of data that corresponds to the traditional fields that fall under the Humanities. Not only does Digital Humanities involve digitizing documents for archival purposes but it also includes material that is considered to be "born-digital". Born-digital data is data that has been created on a computer and only meant to be viewed on a computer and not as a hardcopy document. With material under this category becoming available, it can be seen that one of the goals of this field is to include the integration of multimedia, metadata and dynamic environments. For these different types of data to become available online there is a process of data preparation that must first be completed. This particular study reports on the observations gathered on two real life projects being implemented at the University of North Dakota.

Organizations have been set up to categorize materials that are considered to be born-digital and do not have a paper trail to follow [8]. In this respect, an Electronic Literature Directory was created to provide a database for works that are considered born-digital. Currently, there is a community of electronic literature (e-lit) authors who use XML encoding to tag their work and make it searchable online, while identifying the techniques that were used in the creation of their documents [8]. On the other side, information that originated from a printed textual format should be encoded in XML to allow for indexing and online searching capabilities. Organizations working on digitizing archival documents should establish that the Text Encoding Initiative (TEI) standard be followed so that consistency may be maintained among objects that are made public on the Internet. For example, data included in the category of multimedia of Digital Humanities must meet some specific requirements. From the perspective of a literary scholar, any videos that will be streamed on a web page must be turned into an Mp4 file, while if for downloadable audio purposes it must be converted into an Mp3 file. Some universities and organizations have set up Digital Libraries to store their archival documents and make them accessible and searchable online. A software system developed at the University of Michigan, DLXS, and also the Hathi Trust Digital Library project are examples of collaboration among institutions which are preserving their documents in a digital format.

Several studies have been published in how to handle digital documents. Weibel et al developed a mapping concept between printed information and an interactive digital document [7]. Their study focuses on the logical and physical structure of a document and on the process of presenting the document to the user. The study makes use of metadata to propose a model that makes such mapping more effective. However, it does not emphasize the overall data preparation required for a Digital Humanities document. In the same area, Meneguzzi et al describe a strategy of preparing a document for future printing, using mark-up language as a tool to improve the publishing efficiency of the document [4]. While these two studies are relevant to the area of Digital Humanities, they miss most of the properties that must be archived with the document in order to make it searchable,

comparable, analyzable and easy to be investigated for revision tracking or even copyright infringements.

Antonacopoulos et al were more focused on the data gathering process and published a study on the procedures used to transform historical written documents to electronic ones [1]. Their study focuses on the recovery of scanned original texts and does not mention the use of standards for Digital Humanities or how to handle non-written articles.

In 1994, Gaines, in a very long article, described a set of tools used in the distribution of documents to users, including compact disks and the Internet [3]. When referring to the Internet, he introduced the necessary concepts on the use of HTML to make the document accessible. The study, however, does not mention the necessary data preparation or the use of standards in the characterization of the metadata used in the mark-up languages.

Renear et al described the difficulty of producing meaningful XML metadata in the absence of established semantic rules [6]. This study is of importance for anyone interested in Digital Humanities since it covers the historical significance of mark-up languages and their use. Digital Humanities use XML as its basic metadata language, making this new field highly dependent on mark-up language research. In agreement with the Renear study, the data preparation phase in Digital Humanities must be very well defined and needs to follow careful standards.

Furuta describes the characteristics of documents archived in Digital Libraries [2]. While his study focuses on the importance of the digital document, it does not report on the complexity of the document preparation. This paper focuses on describing a real life, step by step, process of converting a piece of media to a Digital Humanities accessible entity. It discusses the data preparation related to two significant Digital Humanities projects: the Nuremberg Trial Transcripts and the University of North Dakota's Writers Conference [9, 10]. In the next sections, a brief description of the steps involving multi-media components precedes a more in-depth view of the metadata used and the work involved in its preparation. The next section, specifically, provides a background on the XML mark-up language and the TEI standards for Digital Humanities. Such material is followed with a description of the two projects being set-up and the preparation steps required. A more in-depth view of the tag assignments and standard compliance verification follows, including information on how much of it is yet to be automated. Finally, a summary of the presented material closes the study.

2 BACKGROUND

In the area of Digital Humanities there are two basic standards that a creator should be familiar with. One is XML and the other is the Text Encoding Initiative (TEI). XML stands for Extensible Markup Language and is much like HTML. One of the main concepts to understand XML is the idea that it was designed to "carry data" or in other words, it may be used to help with the "indexing" of information so that the tagged information is easily searchable through the Internet. That is also the main difference between XML and HTML. HTML was designed to help a web designer specify how he wants the information to be displayed on a webpage. As stated previously, this is not the purpose of XML. On the Web, XML has its own uses where it is superior to HTML. Since Internet browsers will work with different databases, all at once, the client software is expected to do more of the processing work, and the rendered document will be displayed differently depending on the user's platform and XML directives [5]. XML is defined to be self-descriptive and has no predefined tags. Tags in XML will most often have an open tag and a closing tag. Therefore, even though the creator of the XML document may name the tag as he pleases, he must make sure that he closes the tag. For example, having a tag named "people" would require an opening tag such as <people> and the closing tag would be </people>. Between these two tags is the information which the developer wishes to categorize as "people". A simple example of an XML piece of code is shown in Figure 1.

```
<?xml version 1.0 ?>
<sentence>This is an XML sentence</sentence>
```

Figure 1: XML Code example

The first line in Figure 1 specifies the version that the XML document complies with, the second line uses an element that contains a statement. The element itself is the tag that the programmer placed in the code, in this case <sentence>, and it describes the information that is held between the opening tag and the closing tag.

To be able to write an XML document, the developer must make sure that two objects exist and are compatible. One is a Document Type Defiontiion (DTD) or schema and the other is the document itself. The XML document itself is easy to produce as the developer would just tag the appropriate information. However, in order to design a schema or DTD there are two steps that must be completed. The developer must have a description of the elements and attributes that will be used within the DTD or schema, conducting some research to see if any of it may be reused from a previous XML document. The second step demands the developer to be aware of which

requirements must be met so that the proper element tags are created [5]. However, since the creator of an XML file has all of the freedom to define his own tags, the TEI standards became necessary.

TEI is a consortium which develops and maintains a standard for the representation of data in digital form. These are guidelines that specify the encoding methods for texts, mainly in the area of the Humanities. It is a way to try to ensure that the information posted online will be set up in the same format by all creators who are aware of this set of guidelines. Since 1994, such guidelines have been used by libraries, museums, and publishers. It has been also applied by scholars to present their research and teaching materials. Most significantly, TEI guidelines have served the purpose of preservation of historical or present documents, which has led to the creation of Digital Libraries [12].

3 DATA PREPARATION

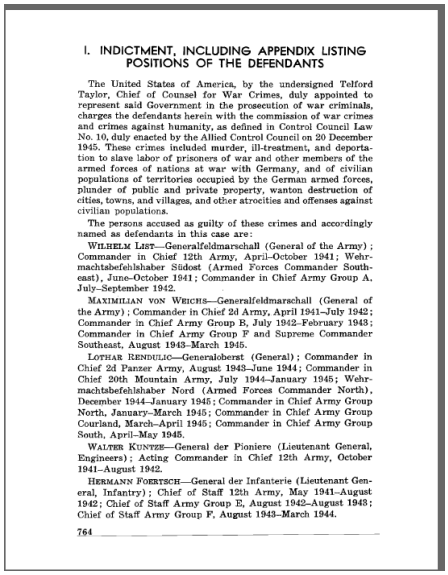
Preparation of digital documents, according to existing standards, has been conducted by several organizations. In this study, we look at two specific projects. One digitization project is being conducted at the University of North Dakota and involves an annual Writers Conference. The University of North Dakota has served as host to this event or more than 40 years. In this conference, well-known authors, poets, prose writers, and playwrights share their work and insights in different sessions. The organizers of the Conference decided on the fourth year of the conference to start videotaping each session. The conference started in 1970 and videotaping began in 1974. By videotaping these sessions, literary and historical documentation were being created. Such documents are currently preserved in the UND Chester Fritz Library.

The UND English Department and the Chester Fritz Library undertook the Writers Conference Digital Project to make the video documentation more widely available. Such project involves converting DVDs and older video media into digital form and offering the conference proceedings through the Internet. The main goal set forth in this project is to convert and make available all sessions through the Library's Digital Collection [11].

A second project involved the Nuremberg Trial Transcripts, held at the University of North Dakota as one of twenty two sets located in the United States, and one of the most complete sets. This collection includes over 240,000 pages of documents archived in the Special Collections at the UND Chester Fritz Library. It not only includes over 300 pages from "the Hostage Case," concerning the Nazi invasion and occupation of Norway from April 9, 1940 to 1944, but also many documents from Justice James Morris, former North Dakota Supreme Court Judge, who presided over the "I. G. Farben Case." The collection was secured in 1949 through the efforts of Dr. Howard Russell, Secretary General of the American Military Tribunals from May 10, 1948 through their completion in December 1949. Dr. Russell was a Professor of English Language and Literature at UND before entering into government services during World War II.

The University of North Dakota's two digitization projects under implementation are, therefore, the Nuremberg Trial Transcripts and the UND Writers Conference. The preparation of the information for digitization in both projects is very similar. However, these two projects differ in the manner in which their documents are provided to those who are working on their digitization. The Nuremberg Trial Transcripts are the actual pages from the trials that took place. The UND Writers Conference data sources are DVD's and older video media that were created of the lecture series that took place during the Conference.

Data preparation for the Nuremberg Trial Transcripts starts by someone having the actual trial transcript in front of them and scanning the document to make an electronic image of it, a process similar to the one described by Antonacopoulos [1] . The documents are scanned at 24 bit color or at 8 bit gray scale. They all must be 5000 pixels in long dimension and saved as a TIFF file without compression. When the scanning is completed, it is considered a "master image" for the project. Figure 2, shown below, is an example of a "master image" that was placed online.



I. INDICTMENT, INCLUDING APPENDIX LISTING POSITIONS OF THE DEFENDANTS

The United States of America, by the undersigned Telford Taylor, Chief of Counsel for War Crimes, duly appointed to represent said Government in the prosecution of war criminals, charges the defendants herein with the commission of war crimes and crimes against humanity, as defined in Control Council Law No. 10, duly enacted by the Allied Control Council on 20 December 1945. These crimes included murder, ill-treatment, and deportation to slave labor of prisoners of war and other members of the armed forces of nations at war with Germany, and of civilian populations of territories occupied by the German armed forces, plunder of public and private property, wanton destruction of cities, towns, and villages, and other atrocities and offenses against civilian populations.

The persons accused as guilty of these crimes and accordingly named as defendants in this case are:

WILHELM LIST—Generalfeldmarschall (General of the Army); Commander in Chief 12th Army, April–October 1941; Wehrmachtsbefehlshaber Südost (Armed Forces Commander Southeast), June–October 1941; Commander in Chief Army Group A, July–September 1942.

MAXIMILIAN VON WEICHS—Generalfeldmarschall (General of the Army); Commander in Chief 2d Army, April 1941–July 1942; Commander in Chief Army Group B, July 1942–February 1943; Commander in Chief Army Group F and Supreme Commander Southeast, August 1943–March 1945.

LOTHAR RENDULIC—Generaloberst (General); Commander in Chief 2d Panzer Army, August 1943–June 1944; Commander in Chief 20th Mountain Army, July 1944–January 1945; Wehrmachtsbefehlshaber Nord (Armed Forces Commander North), December 1944–January 1945; Commander in Chief Army Group North, January–March 1945; Commander in Chief Army Group Courland, March–April 1945; Commander in Chief Army Group South, April–May 1945.

WALTER KUNTZE—General der Pioniere (Lieutenant General, Engineers); Acting Commander in Chief 12th Army, October 1941–August 1942.

HERMANN FOERTSCH—General der Infanterie (Lieutenant General, Infantry); Chief of Staff 12th Army, May 1941–August 1942; Chief of Staff Army Group E, August 1942–August 1943; Chief of Staff Army Group F, August 1943–March 1944.

764

Figure 2: Image of Indictment Page from Nuremberg Trial Transcript

Once a collection of master images has been created, they are written to a DVD. At this point, the master images are deleted from the computer in which they were initially saved. Once the images are on DVD, the next step is to convert the images into JPEG format for web delivery. When converting the images, a set pixel height and width must be declared if there is a style sheet involved that has specified dimensions. If there is no style sheet involved, the person performing the conversion should set constraint proportions and resample image options so that the image is not displayed in an extremely large size on the web site.

After the image has been created, the next step is to transcribe them. Due to the ink on the document fading, Optical Character Recognition is not the most efficient way to transcribe these documents. Therefore, all pages are transcribed manually, according to a minimum set of standards listed next:

1. The spelling, grammar and punctuation must match the original document.
2. For the purpose of the project, it is not necessary to create an exact facsimile representation of the page.
3. Any words that are unclear or illegible should be included to the best of the transcriber's interpretation, enclosed in brackets, or noting in brackets that it is illegible.
4. All page numbers must appear as they appear on the original documents.

The steps followed on the transcription of the Nuremberg Trial Transcripts are the same as those of the Writers Conference, which will be described later in this section. For the Writers Conference, the work done in digitizing the data is a bit more time consuming. During the Conference, different sessions are taped and recorded onto a DVD, however, past conferences may be stored on older video media and must also be converted to DVD format, this is done through a vendor that specializes in preservation and digitization of archival video. For the digitization project, someone takes a DVD (or a .DV file received from the vendor) and converts them to both Mp3 and Mp4 files in order to make them deliverable online. Once the DVD has been converted, a transcriber must watch the conference session and transcribe it. Just as the Nuremberg Trial Transcripts files, the Writers Conference videos are transcribed manually. Based on specifications set up by the Project Manager, all spoken words must be included in the transcriptions; however pauses and voiced pauses, such as "hmmm", do not need to be included unless considered significant to the spoken session.

Once documents and videos have been transcribed, a second person reviews them as a verification step. If there are any issues or disagreements, those are brought up to the Project Manager to make a final decision on the issue. The document is then saved as a Rich Text Format file and converted into an XHTML document. Once the file has been saved, it is then modified to be compliant with the TEI standards. This modification includes updating the header information and using a software program, such as Oxygen, to debug any misplaced XML tags that do not correspond with the standards. Figure 3 shows how a document that has been transcribed and properly tagged in XML up to par with TEI standards looks online. The information shown in Figure 3 corresponds to the transcript page shown in Figure 2. Figure 4 displays what the XML encoding looks like and it corresponds to Figures 2 and 3.

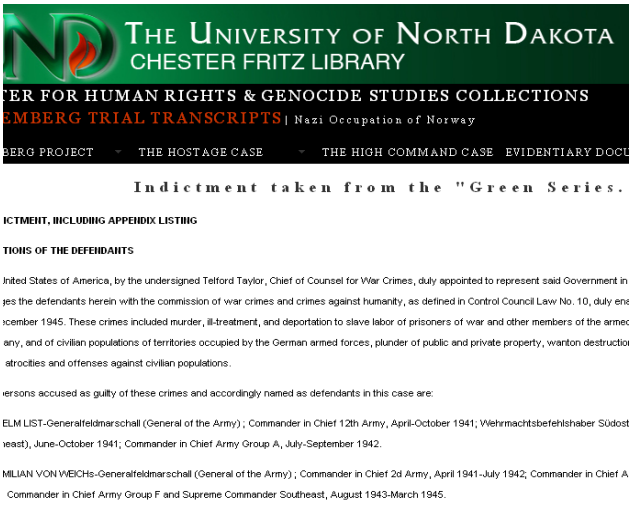


Figure 3: Image of Online transcribed document

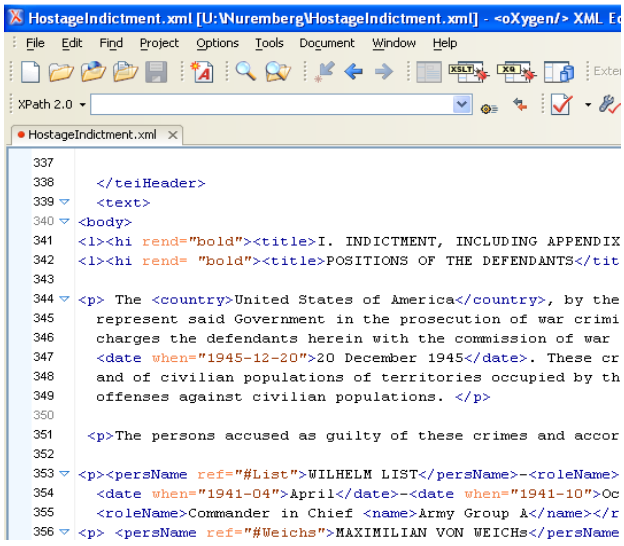


Figure 4: XML Encoding of the Hostage Indictment Transcript

The Writers Conference project is one of the newest projects being implemented. It involves the digitization of approximately 450 hours of footage created during over thirty years of conferences. The project itself has four main goals, first is to be able to archive video from obsolete mediums for preservation purposes, second is to make footage, readings and discussions available over the Internet, third is to transcribe presentations and panel discussions and lastly is to encode these transcriptions, making them TEI compliant by using XML, so they become available online as fully searchable e-texts.

A first prototype was completed based on the 2003 Conference, after permission from all participants was obtained for the audio, video and transcripts to be available online. The project created digital transcripts of the presentations and panel discussions, which complement the streaming video.

During the data preparation, as Mp4 files become available, using video editing software, project staff insert an opening title card and watermark to each file with copyright information as agreed upon by the participants in the video. The watermark, streaming video format, as well as the inclusion of particular java scripts are to protect participant's intellectual property rights and to prevent unauthorized downloads.

To facilitate the transcription process, an open source application, MPEG Streamclip, is used to export the soundtrack for each file to an Mp3 file. Each file is then transcribed into a text document. These documents are reviewed by the Associate Project Director for accuracy. Upon completing the transcriptions, metadata is created in accordance with the guidelines established for the project [10]. The resulting metadata will in turn be reviewed by the Metadata Coordinator and uploaded, along with the corresponding Mp4. At this point, the video footage is available online.

The Library licenses a software system to manage its digital collections. This system uses a text-based search engine built upon Internet standards and protocols. It is optimized for fast text querying capabilities. This provides great flexibility in metadata support and fast performance for large collections. It supports text searches within or across multiple text-based metadata fields, enabling rich metadata searching within or across collections. By employing cataloging standards, a significant number of controlled vocabulary terms are added to every record. This ensures discovery in a consistent and authoritative manner. The library data is also discoverable via web search engines.

4 XML OBSERVATIONS

Once the transcriptions have been verified and the metadata created, the documents are encoded in TEI compliant XML. The methods of encoding and choice of tags are explicitly documented in the processing guidelines and validation techniques are systematically applied in order to ensure accuracy and consistency of the electronic texts [10]. Both the streaming video and the transcripts are placed freely available online for scholarly, educational, and historical purposes.

After being verified, files are prepared for upload to a suite of tools, which were designed specifically for digital library collections. These were deployed by the University of Michigan's Digital Library Extension Services (DLXS). Included in the DLXS suite is a search engine, XPAT, and a XML-aware search engine that is able to index, search, and retrieve UTF-8 Unicode-encoded text data on the fly for web delivery.

The Writers Conference, in particular, is an ongoing project that will continue for the years to come. Since everything that is being done to it, at this moment in time, is all fairly new, there is no current statistical information on how long it may take a transcription of an entire session of a conference or the conference as a whole. Also, the number of XML tags used within a transcription has not been recorded or investigated. However, the Nuremberg Trial Transcripts is a project that is being updated to be placed into the DLXS collection. Some statistical information on this project is shown in the table below for a sample of the files that makes up the collection of the "Hostage Case".

Table 1: Statistical Information for Nuremberg Trial Transcript Procedures

File Name	Total Tags	pers Name	place Name	org Name	pb (page nbrs)
H1_20	1364	77	51	26	20
H21_41	1598	217	225	0	21
H42_83	2007	154	137	77	42
H4_95	694	49	80	25	12
H96_108	677	46	14	25	13
H109_123	859	46	36	5	15
H2545_2625	4163	55	128	0	81
H2626_2680	3392	97	157	0	55
H2681_2734	4628	117	209	39	54

Table 1 contains the number of tags used in each file, and the number of times that the three most common tags within these files are used. Those tags are <persName>,

<placeName>, and <orgName>. In the last column of the table is also a count of the page number tag <pb n="###">, which is a tag that does not have a corresponding closing tag because it only specifies the page number within the tag itself, so that one can see how many pages each file contains. Note that with the <persName> tag each one of them is unique because it makes reference to a person's names. However, for the purpose of statistics all <persName> tags are being counted without splitting the count by particular people referenced. Another tag that is most often found and is not reported in the table is the paragraph tag <p>, this tag is the one that makes up the majority of the tag count shown above. The numbers of tags reported are only those within the body of the XML file, those used in the heading are not included in the statistical figures. Given the number of tags involved, one can estimate the workload involved in manually assigning such tags. This is currently one of the main challenges in the project, the automation or partial automation of the transcription process and data preparation.

5 SUMMARY

Even though the goals that Digital Humanities is trying to achieve, by preserving documents and making them widely available, are not fairly new, the acceptance of the process and use of standards to achieve these goals is a challenge. As shown in this paper the process of data preparation can be extremely tedious and time consuming. However, it has also been shown that parts of the processes are the same even for different data types. With standards, such as TEI, being put into place there is hope that in the near future, digitizing all this different data will be automated and follow the same standards. This is partially being accomplished at the University of North Dakota. The process has not been fully automated however and the Project Manager has made sure that within each of the projects the correct standards are followed by all those who are working on it. The Writer's Conference at the University of North Dakota is an ongoing project that will continue for years and it will be a digitization project that will continue along with the conference itself making the automation process a basic need to be satisfied in a near future.

6 REFERENCES

[1] Antonacopoulos , A., Karatzas, D., Krawczyk, H., and Wiszniewski, B. "The Lifecycle of a Digital Historical Document: Structure and Content," DocEng '04, pp. 147-154, 2004.

[2] Futura, R. "Defining and Using Structure in Digital Documents," DL '94: Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries, 1994.

[3] Gaines, B. R. "Supporting Collaboration Through Multimedia Digital Document Archives," Proceedings of Canadian Multi-media Conference, 1994.

[4] Meneguzzi, F., Meirelles, L., Mano, F., Oliveira, J., and Silva, A. "Strategies for Document Optimization in Digital Publishing," DocEng '04, pp. 163-170, 2004.

[5] Mercer, D. XML: A Beginner's Guide, McGraw-Hill, Chicago, 2001.

[6] Renear, A., Dubin, A., Sperberg-McQueen, C.M., and Huitfeldt, C. "Towards a Semantics for XML Markup," DocEng '02, 2002.

[7] Weibel, N., Norrie, M., and Signer, B. "A Model for Mapping between Printed and Digital Document Instances," DocEng '07, 2007.

[8] National Endowment for the Humanities, Electronic Literature Directory. http://directory.eliterature.org/. June 29, 2011.

[9] The University of North Dakota Chester Fritz Library, Center for Human Rights & Genocide Studies Collection: Nuremberg Trial Transcripts. http://www.und.edu/instruct/calberts/Nuremberg/NurembergMain.html. June 29, 2011.

[10] The University of North Dakota, Writers Conference. http://www.undwritersconference.org/. June 29, 2011.

[11]The University of North Dakota, Chester Fritz Library. http://library.und.edu/digital/writers-conference/. June 29, 2011.

[12] Text Encoding Initiative. http://www.tei-c.org/index.xml. June 29, 2011.