JDHCS 2010
Volume 1 Number 2

Bringing It All Together: Integrating Text, Audio, Metadata, GIS, and Scholarly Criticism in a Holocaust Oral History Archive

Eben English, Paul V. Galvin Library, Illinois Institute of Technology

Introduction

In July 1946, Dr. David P. Boder, a psychology professor from Illinois Institute of Technology (IIT), traveled to war-torn Europe to record the stories of Holocaust survivors in their own words and in their own voices. Over the next three months, he visited refugee camps, orphanages, and rehabilitation centers in France, Switzerland, Italy, and Germany, carrying with him a portable wire recording device and 200 spools of steel wire, upon which he was able to record over 120 interviews comprising over 100 hours of material in nine different languages. These wire spools, and the narratives they contain, represent the earliest known oral histories of the Holocaust.

Despite the groundbreaking nature of his work, Boder was largely unsuccessful in his efforts to publish the interviews upon his return to the United States. Seeking to preserve his work at the end of his career, he submitted a set of 70 interview transcripts to a select number of libraries and historical foundations across the U.S. (including IIT), though few volumes remain today. Boder died in 1961, leaving countless hours of interview material still to be transcribed and without a permanent institutional home.

The Voices of the Holocaust project,[1] curated by the Paul V. Galvin Library, was born in 1998 with the discovery of a set of transcripts of Dr. Boder's interviews in the IIT University Archives. Since that time, the project's mission has been to digitize, restore, transcribe, and translate Dr. Boder's historic recordings so that they can be experienced online by a global audience of students, researchers, historians, and the general public.

Collections in the Dark

There are numerous archives of Holocaust survivor testimony in existence today. However, very few of these collections are available online, and even fewer use standards-based text-encoding and markup practices, which limits both the possibilities for effective online dissemination and the prospects for long-term preservation. The metadata used to describe the interviews and their participants in these collections often lacks sufficient detail, placing a significant burden on researchers searching for information on the experiences of a particular survivor demographic. In addition, scholars currently have no means to conduct research across disparate collections of survivor testimony maintained by different institutions.

Representing the rich variety of media manifestations produced through the interview process requires more sophisticated data modeling practices than are currently used by most archives. Holocaust survivor interviews are not merely transcripts of recorded conversation accompanied by speaker annotations—they usually have an accompanying audio or video component as well. As Goldman et al. note: "a spoken message contains more than simply what was said and who said it. The prosody—timing, intonation, and stress—of the speech signal offers a great deal of information about the emotional state of the speaker, 'punctuation' in the speech and

[1] http://voices.iit.edu

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

disambiguation of the intended message."[2] Unfortunately, the relationship between these manifestations of the interview event is often obscured in existing collections, meaning that this unspoken information is lost.

Data Modeling Requirements for Holocaust Survivor Testimony

To properly address these challenges, a data model specifically tailored to Holocaust survivor testimony is needed. This model must support the needs of a wide variety of scholars, including historians, psychologists, social scientists, and linguists, each of whom are interested in different semantic and structural aspects of the documents. It must also preserve the integrity of the interview as a holistic unit and make explicit the relationships between the text(s) and audiovisual manifestation(s).

From a more technical standpoint, this data model should support detailed descriptive, administrative, and structural metadata; allow for synchronization of text data with an audio or video bitstream; support multiple languages and character sets; allow for the association of other types of content with the material (e.g. GIS, data sets, critical and contextual content such as glossaries, footnotes, etc.); allow the data to be easily transformed, cross-walked, or migrated; be able to be validated against an established schema with an established community of support; and of course, provide a solid framework for flexible, robust online presentation.

Towards an Implementation

In September 2009, the newly redesigned and revamped Voices of the Holocaust site was launched. At its core, the collection consists of text and audio for 118 interviews with Holocaust survivors and other displaced persons. However, the site provides much more than just static facsimiles of the source documents—it offers a host of other interactive features, making it an invaluable historical resource for scholars and students of all ages.

To accomplish the project's goal of enhancing the delivery of the interviews to make them more interactive, educational, and engaging, and to meet the requirements discussed above, an XML-based data model was created based on the TEI P5 schema, which allows for the integration of many different types of content (transcriptions, audio, metadata, GIS, and scholarly criticism) into a unified presentation for the user.

Site features include original-language transcriptions and English translations of each interview; digitally remastered audio of the interviews with synchronized transcripts; full-text searching of interview content; advanced searching and browsing based on biographical, historical, and geographical facets of the interviews and interviewees' testimony; interactive maps featuring concentration camps, ghettos, interview locations, interviewee birthplaces, and other relevant locations; criticism and commentary from prominent Holocaust scholars, including introductions, footnotes, a glossary of terms, and an extensive bibliography; and detailed technical notes on the project's development, including sample XML files and source code examples.

The collection is made up of four distinct data types: text, audio, metadata, and GIS. Each of these data types, and the means by which they are integrated with one another, are discussed below.

---

[2] Goldman et al.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Text: The interview texts, including original-language transcriptions and English translations, are encoded in XML using the TEI P5 schema. The TEI schema facilitates flexible display, detailed searching, portability, and provides the necessary structure to connect multiple data sets and media types to discrete elements of the text. Nine different languages are represented in the collection, as well as Latin, Cyrillic, and Hebrew character sets (often mixed within the same interview), which are handled by using UTF-8 character encoding. The interview texts are divided and marked up into utterances using the <u> element and a corresponding attribute identifying the speaker name and role (interviewer, interviewee, translator, etc.). This level of atomization allows each utterance to be associated with a specific temporal location within the audio file, and for a separate index of questions and responses to be created. Full-text searching is accomplished using Solr, an open-source enterprise search server.

The presence of scholarly criticism and contextual material in the collection adds immense value to its potential as an educational resource. (Illinois state law mandates genocide and Holocaust education for elementary and high-school students.[3]) By their intrinsic nature, interviews are subjective recollections rather than objective historical narratives, and a critical perspective to guide the reader is essential. Critical content for the site was voluntarily contributed from Holocaust scholars and historians, and includes introductions, footnotes, a glossary of terms, and an extensive bibliography. This material is directly integrated into the TEI XML files for each transcription using the <note> element, which is placed directly into the encoded text at the location where a footnote would appear in a printed version. These notes can be accessed dynamically during reading via JavaScript-enabled pop-ups and rollovers.

Audio: The TEI schema contains provisions for encoding the start and end times of speech elements in a transcription of audiovisual material. Synchronizing text and audio files for simultaneous presentation is accomplished by inserting time-code information (in milliseconds) into the XML files as the value of an attribute of the <u> element containing the text for the corresponding utterance. Time-code information is derived using Transcriber, an open-source tool for segmenting and labeling audio files. For presentation, a Flash application reads the time-code values and highlights the text for the audio being played. This means that the user can listen to a recording in any language and follow along with an English translation, retaining the essential prosodic information.

Metadata: In order to facilitate advanced searching and browsing features, it was necessary to define a metadata schema which could accommodate detailed biographical information specific to Holocaust survivors, such as tattoo number, location at time of Nazi occupation, camps/ghettos interned at, liberation date, etc. These elements fall outside the scope of most common metadata schemas (such as Dublin Core) which are used to describe documents rather than intellectual content. The TEI schema contains provisions for many biographical metadata elements (age, religion, nationality), and others (such as tattoo number) can be accommodated with some creative tweaking. These metadata elements are directly encoded within the <teiHeader> element in the XML documents. Solr is used to index the metadata.

GIS: Interactive maps provide a powerful tool for users to visually browse and interact with the collection. Geographic data—including location names and coordinates—is not only used in

[3] http://www.isbe.state.il.us/news/2005/aug5_05.pdf.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

interactive maps, but also inserted directly into the interview texts to give the user an in-context reference source as they read the interview. GIS information is maintained in a separate master file (also in TEI XML format) to ensure referential integrity, and incorporated into the interview text files using XInclude. References to specific locations in the interview text are marked using the <placeName> element, which takes as an attribute the identifier of the location in the master file. The interactive maps are created using JavaScript, OpenLayers, and Google Maps applications.

Conclusions & Future Directions

The data model developed for the Voices of the Holocaust collection represents a significant step towards meeting the research community's need for powerful searching of, dynamic access to, and long-term preservation of Holocaust survivor testimony, most of which is still hidden inside archives and unavailable online. Using XML-based data modeling and open source technologies, the Voices site combines interview transcriptions with their accompanying recordings, integrates GIS and scholarly content to provide an enriched user experience, and offers robust searching and browsing functionality based on historical, biographical, and geographical metadata tailored to the subject area. It brings the survivors' voices to life in a way never before possible, increasing the visibility and impact of the interviews, promoting deeper scholarship and analysis, and ultimately providing a richer understanding of the Holocaust and those who experienced it.

Now that an example of such a data model has been implemented, it needs to be reviewed and revised by a community of scholars, archivists, librarians, and data curators, and put into practice by other projects and collections. In addition to making research more efficient, widespread adoption of this model could lead to some very exciting opportunities, including increased discoverability of testimony content through search engines, integration with other types of oral history collections, cross-searching of multiple collections simultaneously, and the development of automated linguistic or content analysis tools.

Acknowledgments

Funding for this project was awarded by the Illinois State Library (ISL), a Department of the Office of Secretary of State, using funds provided by the U.S. Institute of Museum and Library Services (IMLS), under the federal Library Services and Technology Act (LSTA).

References

Goldman, J. et al. 2005. Accessing the spoken word. *International Journal on Digital Libraries* 5(4): 291.

Office of the Illinois Governor. 2005. Gov. Blagojevich signs law expanding genocide education in Illinois. http://www.isbe.state.il.us/news/2005/aug5_05.pdf. (accessed November 11, 2009).