

Final Report on “XML encoded finding aids of Holocaust survivor testimonies”: SCOPA grant award winner for FY 2003/2004

Background:

The Fortunoff Video Archive for Holocaust Testimonies (VAHT) currently holds 4,267 videotaped testimonies of Holocaust witnesses, which comprise over 10,000 hours recorded hours of videotape. This number has grown from slightly fewer than 200 when VAHT became a full-program of MSSA in 1982.

In terms of providing intellectual access to these materials there were several false starts, along with some personnel changes. When MSSA became one of the founders of RLIN-AMC, it was decided that the bibliographic records would be entered into RLIN for each testimony. The initial method of cataloging was based on the archivist listening to the testimonies and taking handwritten notes.

In 1995, we began creating these notes on PCs in a word processing program. We are in the process of converting the older handwritten finding aids to word documents. At the present time we have completed notes for 3,395 testimonies, in a mix of handwritten and electronic forms. We hope that the conversion process from the handwritten notes will be completed within the next eighteen months; however, creation of notes for those testimonies to which students have not yet been assigned will take considerably longer, particularly since many are in Hebrew, French, Czech, and Yiddish.

Included in these notes are timestamps that indicate at what time during the videotaped testimony a "speech event" takes place. In this sense the testimony notes serve as a "finding aid", permitting the researcher to identify the section of videotape that is of most interest to her. Over the years, as staff have read these finding aids, manuscript annotations, such as corrections, notes and explanations have been added, and these annotations themselves inform and increase the utility of the finding aid. Such annotations, however, written with many hands, over time, are often hard to decipher. The wide variation in the printed finding aids may be an impediment to researchers effectively finding the information they require.

As initially stated the project had six primary goals, as outlined in the original proposal in **Appendix A**. Over the following pages we shall comment on the progress toward each of these goals, providing detail on the challenges faced, and how and if they were overcome.

Goal 1: Encode approximately 200 finding aids

Encoding was envisioned as taking place on three levels, using three markup techniques to create a completed TEI-encoded finding aid in XML, from which we could then produce HTML for web display and PDF for standard print formatting.

Firstly scripts would be written to create “base TEI” from finding aids currently in Microsoft Word. These finding aids have been created over a number of years by dozens of undergraduate and graduate students. Although they were given instruction in how to create the finding aids from watching the video-recorded testimony, inevitably variation in how they have chosen to interpret those rules have led to finding aids of great variety. Compounding this, as the students

are working on DOS workstations, without mice or a graphical user interface, using an old version of WordPerfect, we find imperfections borne both of a lack of familiarity with this older system, and inaccuracies in the subsequent migration to Word.

There are certain hooks, however, that have allowed us to create scripts that do a remarkably effective job in “flattening” this variation, and outputting XML. The XML at this point is not fully compliant TEI, but it does meet our requirement for **“level 1” markup**. At this point the XML can be indexed and retrieved from a local prototype system.

Once we have satisfactory (i.e. well-formed XML) level 1 markup, the XML files are passed to students to further process the files to **“level 2 markup”** status. Given a freely available XML editing application, an XML file, a printed copy of the finding aid in question, and an instructional document, a student proceeds to encode manuscript annotations on the finding aid. The annotations are corrections, elaborations and other general editorial comments on the quality and accuracy of the initial finding aid, by a qualified archivist in the Fortunoff Video Archive for Holocaust Testimonies.

We currently have 263 XML files at “level 2 markup”, and have spent \$273.58 on student wages. We also purchased a software license for an XML editing program (XmlShell) for \$45.00. We hope to encode approximately 400 more finding aids using the remaining funds..

“Level 3 markup” was the final aim of encoding, and consist of two portions:

1. generating subject headings from MARC records held in Obis for the corresponding video-cassette, and including these programmatically in the TEI header
2. creating authoritative terminological thesauri for markup, e.g. place names

Step one has been technically proven, but along with step two there has been a policy decision to wait on level three encoding until we have a critical mass (perhaps 75% of the total) on finding aids encoded in TEI. One major factor in holding off on both phases of level 3markup has been that many MARC records remain provisional or suppressed at this time.

Goal 2: Implement a flexible search interface to a native XML database (access may be restricted owing to the sensitivity of certain testimonies)

A database has been built, one based on the SWISH-E indexing software <http://swish-e.org/>, and local PHP code. But how this database is used given the sensitive nature of the finding aids is still up for discussion. **Figure 1** shows the result set from a search for “holocaust” on a development version of the database, one that also indexes EAD and EAC files, and XMLized MARC records.

Technically we can deliver the finding aids from the database in a format that very much mirrors the display of our EAD files (one benefit of going to TEI), but we are not sure how the access mechanism will finally pan out. It may well be that the effort of developing the database and cross-XML searching is not sufficient to warrant the problems inherent in having such materials online.

Several privacy issues will prevent MSSA patrons from enjoying unhindered access to encoded Holocaust testimony finding aids. The encoded finding aids, with marked-up editorial comments, must therefore be kept under some form of password protection. After registering in the reference room, patrons will receive a password (changed daily?) that allows them to access the electronic finding aids through MSSA's finding aid database. Every testimony with a corresponding electronic finding aid will be listed in the database. These listings will be similar to the MARC records currently available in Yale's Orbis catalog. The main difference, however, is that the MARC-like records in the finding aid database will include a link to the full text of the finding aid. This link, as mentioned above, will not activate without the password the patron receives during registration. The electronic finding aids available to patrons must not be printable, and patrons should not have the option of cutting, pasting and sending portions of the finding aid in an email document to themselves or another person. The personal names and other sensitive information contained in these finding aids must be protected, even if this creates an impediment to the realization of immediate, open access, which often lies at the heart of such electronic initiatives.

Finding Aid Listing	
1	<p><u>Personal Holocaust Testimony of</u> personal names removed</p> <p>Relevance: 1000 Size: 12 KB File type: TEI.2</p>
2	<p><u>Personal Holocaust Testimony of</u> Personal names removed</p> <p>Relevance: 944 Size: 15 KB File type: TEI.2</p>
3	<p><u>Personal Holocaust Testimony of</u> Personal names removed</p> <p>Relevance: 922 Size: 30 KB File type: TEI.2</p>
4	<p><u>Personal Holocaust Testimony of</u> Personal names removed</p> <p>Relevance: 911 Size: 15 KB</p>

Figure 1. Result set from searching TEI-only in the local database

Goal 3: automatically generate EAD encoded finding aid from the encoded finding aids

The feasibility of this has been tested only conceptually and no development work has been conducted in providing an actual mechanism for accomplishing this at this time. The limiting factor, once again, has been the desire to push back until a significant proportion of the finding aids have been encoded. In the case of the production of a single EAD finding aid, this proportion will have to represent 100% of existing finding aids. EAD is, of course, an encoding standard to describe a collection of materials. At this point we feel that an EAD (*viz. collection level* finding aid) that is incomplete would be misleading to our patrons.

Goal 4: create HTML and PDF versions of the finding aids programmatically

Appendix B illustrates a portion of a PDF file created directly from a TEI encoded XML document instance. The PDFs are created by the application of XSLT (eXtensible Stylesheets for Transformations) producing XSLFO (eXtensible Stylesheets for Formatting Objects) which is then rendered to PDF using Apache's FOP (Formatting Object Processor), <<http://xml.apache.org/fop>>.

By using XSLFO as our pre-delivery intermediary, we are also able to create Rich Text Format files for editing, ASCII text files for cross-system transparency, or even the emerging Scaleable Vector Graphic Format (SVG).

Goal 5: create a sustainable set of tools so that future finding aids may produce TEI encoded documents

A provisional template (vide **Figure 2**), subject to student user testing, has been created in Open Office the open source rival to Microsoft Office, from Sun Microsystems <<http://www.openoffice.org>>. Open Office files are actually zipped collections of XML files containing a native XML file format, a manifest log, some rendering information and some general technical metadata. Each XML file in the zipped package is read by Open Office to produce an environment in which one can create and edit documents.

Following a model developed for the creating of EAD finding aids, we shall develop scripts that unpack the zip archive, read the contents of the native XML files and convert those to TEI. Portions of the scripts and their operational pipeline have been completed, but we are awaiting the finalization of the format, and this is hindered solely by hardware.

Students currently creating new finding aids are doing so on mouse-less, GUI-less machines; come the next equipment request cycle, we will be able to deploy the template.

Student note-takers create testimony notes or finding aids according to a set of guidelines (vide **Appendix C**). Nevertheless, finding aids created by different students vary greatly. This template, in turn, will be scripted for quick conversion to XML. The template will contain three categories of information in addition to the "transcriptions" of speech of the interviewer(s) and testifier(s):

A. Information input by students:

HVT number
Testifier(s)' name
Date of recording
Time
Interviewer(s)' name
Name of student note-taker
The actual text of the testimony

B. Information input by the archivist

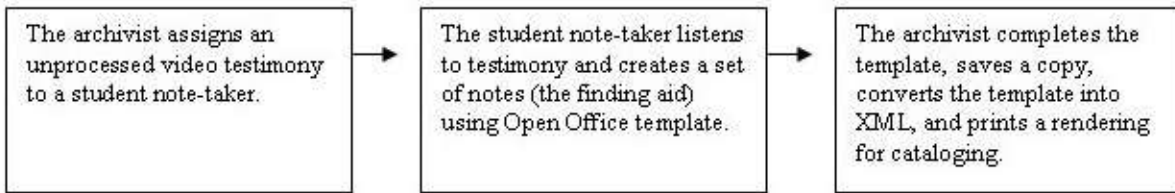
Edited by
Equipment description, i.e. PAL or NTSC
Project description (affiliate project, etc.)
Testifier's gender
Testifier's date of birth
Restrictions
Language in which interview was conducted

C. "Boilerplate" information that will populate automatically

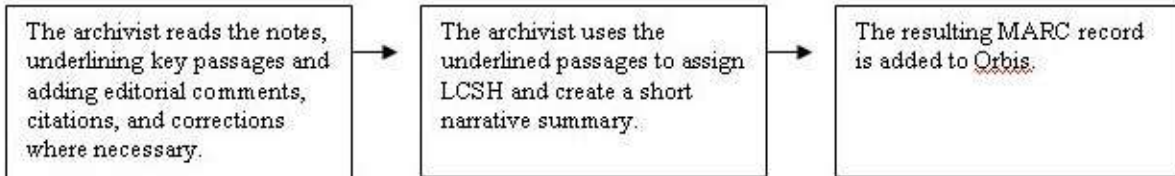
Sponsor information
Principal investigator information
Publication information
Availability information
Caveats concerning the production of the text
Admonishments not to rely on the notes, but to always check this against the use-copy of the video-cassette.
And generic statements as the nature of the production of the text

Since essentially the same model for creating XML encoded texts in already in production mode in Manuscripts and Archives we are able to already project a workflow scenario, which anticipate to take the following outline:

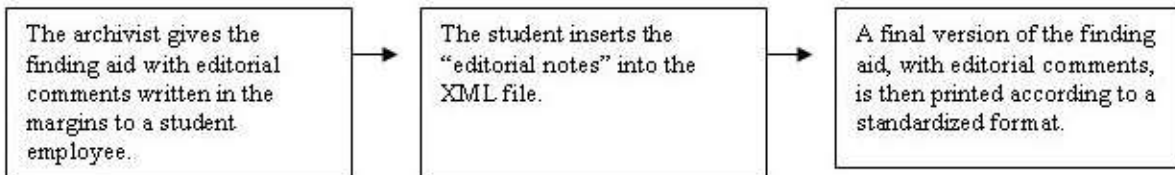
Step 1: Creating the finding aid



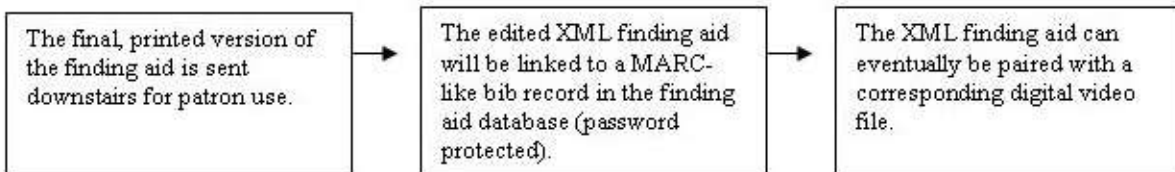
Step 2: Cataloging



Step 3: XML



Step 4: Delivery





OO2TEI 0.2: Template for VAHT Holocaust Witness Finding Aids	
<i>Testimony #</i>	T-0000
<i>Interviewee</i>	FirstName SecondName [multiple values separated by a hard return]
<i>Interviewer(s)</i>	FirstName SecondName [multiple values separated by a hard return]
<i>Interview date(s)</i>	Month dd, yyyy [multiple values separated by a hard return]
<i>Note taker</i>	FirstName SecondName
<i>Notes Date</i>	[AUTODATE]

Figure 2: Screenshot of the OpenOffice Template

Conclusion

Our broad vision in outlining the scope of our initial proposal rather, that is, that the deliverables has encompassed three areas: 1. technical innovation; 2. convergence in look-and-feel; and 3. improved workflow process. In each of these three areas we have made some notable successes, and these measures we think that those who have been involved in the project have made important contributions to the department, and hence the Library.

This award of this grant has allowed us space to develop ideas and tools for cross-searching textual resources, and has created in the department an ongoing program to encode witness testimonies in a structured manner.

Special thanks go to Raman Prasad and Stephen Naron for their hard work in Microsoft Word document parsing and Open Office template design respectively.

Joanne Rudof
Stephen Yearl

May 2004

Appendix A: The Original Proposal

XML encoded finding aids of Holocaust survivor testimonies

Background:

The Fortunoff Video Archive for Holocaust Testimonies (VAHT) holds more than 4,100 testimonies of Holocaust survivors, which are comprised of over 10,000 recorded hours of videotape. Transcriptions of many of the videos have been made, and are in various word processor formats, printed copies of which are made available to researchers. Included in these documents are timestamps that indicate at what time during the videotaped testimony a "speech event" takes place. In this sense the testimony serves as a "finding aid", permitting the researcher to identify the section of videotape that is of most interest to her. Over the years, as scholars and staff have read these finding aids, manuscript annotations, such as corrections, notes and explanations have been added, and these annotations themselves inform and increase the utility of the finding aid. Such annotations, however, written with many hands, over time, are often hard to decipher. The wide variation in the printed finding aids is an impediment to researchers effectively finding the information they require.

Proposal:

It is intended that we will develop or adopt an XML encoding scheme to structure the finding aids and their annotations in order to facilitate access and retrieval, and further to process them uniformly for consistent display in print and on screen. Due to the oftentimes sensitive nature of information in the finding aids, we will also implement access and restriction policies, possibly based on Internet Protocol Address. We envision that the finding aids be encoded using the Text Encoding for Interchange (TEI), spoken language section.

Further, we hope to index these finding aids in a database that will allow searching on specific XML "fields". Database results will, additionally, be converted to HTML on-the-fly, using server-side scripting. To complement the HTML, we aim also to create consistently formatted PDF files for print.

Given a significant number of TEI encoded finding aids, we hope to generate an EAD finding aid to serve as a Guide to the Holdings of the archive, and possibly contribute this EAD instance to the Yale Finding Aid Database.

Aim to:

- + Encode approximately 200 finding aids
- + Generate significant portions of testimony metadata, by script, from MARC records in Orbis
- + Implement a flexible search interface to a native XML database (access may be restricted owing to the sensitivity of certain testimonies)

- + automatically generate EAD encoded finding aid from the encoded finding aids
- + create HTML and PDF versions of the finding aids programmatically
- + create a sustainable set of tools so that future finding aids may produce TEI encoded documents.

benefits to department:

- + uniform formatting of all VAHT testimonies, providing a consistent rendering for readers
- + possibility of integrating with the Yale EAD finding aid database
- + explore the possibility of programmatically creating MARC records from the TEI
- + enhanced transcript management functions, such as ordering/sorting finding aid by, for example, name, transcript number, geographical place, etc.
- + ability to extract data for addition to VAHT/MSSA databases

wider benefits:

- + formatted finding aids will look like other printed archival finding aids, such as those at BRBL or MSSA
- + evaluation of a native XML database system that allows complex searching on specific "tags"
- + methodology for dynamic display of XML documents "on-the-fly"
- + development of an encoded speech transcription scheme that might usefully be adopted by others, such as encoding oral histories, lectures, or other oral performances.

Request:

Since we intend to use open source software, and the department has adequate hardware for the storage and delivery of the encoded descriptions we are asking only student time to TEI encode the finding aids. Preliminary estimates indicate an XML encoding rate of 2 finding aids per hour on average.

ca. \$1000 for student labor

according to:

200 finding aids (2 p/h) = 100 hours = \$1000 (estimating \$10 p/h for student labor)

Appendix B: Comparing EAD and TEI PDFs

This screen capture illustrates one of the primary stated goals of the project: to bring into closer alignment the representation of EAD and TEI based finding aids into a common look-and-feel in order to reduce patron confusion and increase internal consistency.

The screenshot shows a web interface with two main panels. On the left is the 'Overview of the Papers' panel, which is circled in blue and labeled 'PDF of EAD'. On the right is the 'Transcription Overview' panel, which is circled in orange and labeled 'PDF of TEI'. The interface also includes an Acrobat Reader window at the top and a central navigation pane with 'Bookmarks' and 'Thumbnails' tabs.

Overview of the Papers	
CREATOR	DeMare, Merc
TITLE	Mercedes Moo
PHYSICAL DESCRIPTION	1 linear ft.
ARRANGEMENT	Arranged by ty
Biographical Sketch	Mercedes Moo Brooklyn, New and received a Arts in 1936. 1 Broadway prod
SUMMARY	The papers con materials belo graduate of the Fine Arts.
PROVENANCE	Gift of the Woc
TERMS GOVERNING USE	Copyright has s
PREFERRED CITATION	Mercedes Moo University Libr
FOR FURTHER INFORMATION	Manuscripts an Yale University P.O. Box 2082 New Haven, C

Transcription Overview	
TESTIFIED	personal names removed
INTERVIEWERS	Bernard Weinstein, and Freda Remmers.
DATE RECORDED	October 29, 1987
DATE TRANSCRIBED	October 2, 1999
TRANSCRIBED BY	Transcribed by Cristina Sloan
TERMS GOVERNING USE	Copyright 1993-2002, Yale University Library Available only with prior consent of the Fortunoff Video Archi for Holocaust Testimonies.
PREFERRED CITATION	"Personal Holocaust Testimony of [personal names removed] (T-1179)," at T Fortunoff Video Archive for Holocaust Testimonies, Manuscrip and Archives, Yale University Library.
FOR FURTHER INFORMATION	Fortunoff Video Archive for Holocaust Testimonies Manuscripts and Archives Yale University Library P.O. Box 208240 New Haven, CT 06520-8240

Appendix C: Guidelines for Note takers

Fortunoff Video Archive for Holocaust Testimonies Guidelines for taking testimony notes

1. At the beginning of each set of notes please include the T-number, the name of the witness, the date of the interview, and your name. If available, please add the name of the affiliate project and the names of the interviewers. Mark the running time of the testimony at least once every five minutes, using the elapsed time clock on-screen. Place the time code on the left margin on its own line.
2. There are no hard and fast rules about how long a testimony will last. Some are barely 30 minutes; others are 30 hours. A typical testimony may last two hours, with a tape break after about an hour. (Sometimes tape breaks occur every twenty minutes).
3. You are making detailed notes, not a transcript. It is vital to capture the narrative and all place names and dates that pertain to the witness' own story. Do not spend time looking up information which is hearsay - that is, about a third party. Please verify all geographic names pertaining to the witness' story. Obviously familiar places (e.g. Warsaw, Paris, etc.) do not need verification. Unfamiliar places should all be checked in the reference books. Please note parenthetically the source and page number you used to verify the place name. **DO NOT GUESS!!!** If you cannot understand a word after checking the reference books and glossaries, record the word phonetically in brackets and note the precise time (hour, minute, second) when it was said. Also record this information on a piece of paper so when you consult Video Archive staff, you do not have to go back to your disk. Leave time to discuss these words and other questions with Video Archive staff at the end of each session. It is very important to bring questions to us as you encounter them: you will be able to place the word in context for the staff, and quickly locate its place on the tape. If we resolve such questions immediately, it saves an enormous amount of staff time and effort later on.
4. Please use the first person, e.g. "I was born in Paris...."; otherwise, the many third-person pronouns may be confusing to the reader. While it is often difficult to impose a logical paragraph structure on a testimony, occasional line breaks and indented first lines will increase the readability of your work enormously. For similar reasons, please use ordinary rules of capitalization and punctuation.
5. Please make a note of any problems with the testimony: if the video or audio seems to be damaged or of poor quality; if there seems to be segments of testimony missing or out of sequence; or if after checking with the Video Archive staff, you find that a segment of the interview is too confusing to understand.
6. The listed reference books (see below) are available in the Reading Room for consultation. Other references are shelved in the Video Archive and are available for consultation either before or after a session in the Reading Room. Most testimonies will have a pre-interview form, which should be on your shelf with the cassette. (If it isn't, please ask). These forms list names and places that may be mentioned in the interview; please remember that pre-interview forms, while

extremely useful, are not considered authoritative sources. All the information they contain needs to be verified in the available reference works.

7. To verify camp names, check the references in this order:

a. *Die nationalsozialistische Lagersystem* (thick grey book; the Video Archive staff refer to this as the *ITS*; it was published by the International Tracing Service of the Red Cross)

b. Martin Gilbert, *Atlas of the Holocaust*

8. To verify the names of cities and towns:

a. *New York Times Atlas*

b. *Where Once We Walked*

c. Gilbert, *Atlas*

9. The *Encyclopedia Judaica* and *Encyclopedia of the Holocaust* are both available in the Video Archive. They can be used as references either at the beginning or end of each work session. Questions that arise during each session may require these references or Video Archive staff assistance. This should be recorded and addressed at the end of each session.