Survivors of the Shoah Visual History Foundation: an introduction to its indexing methodology

The Shoah Visual History Foundation in cooperation with Mary Crystal

Under the guidance of leading scholars and historians the Survivors of the Shoah Visual History Foundation has videotaped the testimonies of Holocaust survivors worldwide in 29 languages. Using state-of-the-art technology a unique indexing system was created to facilitate access to the place names, personal names, events, experiences and other topics in the nearly 50,000 interviews conducted thus far. As studies in indexing multi-media content move forward, the Foundation hopes to create a research center for advancing indexing in the area of digital asset management. This article provides an introduction to the Foundation’s unique indexing methodology which has been developed to enable the delivery of survivors’ oral testimonies.

Background: the holocaust
The Holocaust was the systematic, bureaucratic annihilation of six million Jews by the Nazi regime and its collaborators as a central act of state during World War II. In 1933 approximately nine million Jews lived in the 21 countries of Europe that would be occupied by Germany during the war. By 1945, two out of every three European Jews had been killed. Although Jews were the primary victims, hundreds of thousands of Roma and Sinti (Gypsies) and at least 250,000 mentally or physically disabled persons were also victims of Nazi genocide.

As Nazi tyranny spread across Europe from 1933 to 1945, millions of other innocent people were persecuted and murdered. More than three million Soviet prisoners of war were killed because of their nationality. Poles, as well as other Slavs, were targeted for slave labor, and as a result, almost two million perished. Homosexuals and others deemed “anti-social” were also persecuted and often murdered. In addition, thousands of political and religious dissidents such as communists, socialists, trade unionists, and Jehovah’s Witnesses were persecuted for their beliefs and behavior and many of these individuals died as a result of maltreatment.

Shoah, a Hebrew word meaning “desolation,” has come to be the preferred term for the Holocaust by Jewish scholars.

Background: the Shoah Foundation
The Survivors of the Shoah Visual History Foundation was founded in 1994 by director Steven Spielberg following the release of his award winning film Schindler’s List. The nonprofit organization is dedicated to videotaping and archiving interviews of Holocaust survivors all over the world. These testimonies are digitized and indexed, and will be made available via on-line, interactive networks to museums, educational institutions and nonprofit organizations as well as through educational documentaries, books and on CD-ROMS.

The Foundation software suites enable the cataloging and distribution of hundreds of tera-bytes of multi-media data including digital video, stills, and biographical information. These software suites allow queries such as “tell me about children who received education in the Warsaw Ghetto” to return rich multi-media data sets consisting of pertinent information from the testimonies themselves. The indexing system enables end users to arrive at segments, i.e. short sections of testimony which are relevant to the query.

Once indexed, the MPEG versions of the testimonies and the associated metadata are made available through media delivery software written at the Foundation which utilizes products incorporating standard CORBA, RDBMS and ATM to deliver the media to an initial five Holocaust museums and libraries. Organizations such as EMC, GTE, SGI and Sybase have supplied the enabling technology for the Foundation’s digital library system.

Collecting and processing the testimonies
The process for collecting a testimony starts with training interviewers and videographers. Prior to the interviews, participating survivors are pre-interviewed using a forty-page document to collect pertinent background information. Conducted in the language most comfortable for the survivor, interviews are taped in the survivor’s home (unless another location is requested) using broadcast quality Betacam SP video equipment. The tapes of the interview and the pre-interview questionnaire are then shipped to the Foundation.

Once the tapes are delivered to the Foundation they are then
digitized into MPEG and stored on a 150 tera-byte tape archive. From here the testimonies are indexed and prepared for distribution. The initial distribution network (Figure 1) currently includes four sites in the continental United States and one site in Israel.

**Thesaurus construction**

The Foundation's cataloguing department organizes index terms through its specially developed thesaurus. Use of this controlled vocabulary alleviates the following problems inherent in querying data.

1. It allows end-users to find appropriate terminology.
2. It enables end users to conduct more powerful searches.
3. It ensures the consistency of terminology used to describe the data.

Development of the thesaurus incorporated both the selection of appropriate index terms and the establishment of relationships between them. At this point the cataloging department is at the early stages of applying the thesaurus to the collected interviews in order to create an archive of Holocaust survivor and witness testimonies.

The structure of the thesaurus is based on the ANSI/NISO Z39.19, *Guidelines for the construction, format, and management of monolingual thesauri*. When it created the thesaurus the cataloging department established a core of index terms and also established a process for absorbing additional terms as indexing continues.

When the development of the thesaurus commenced, the Foundation solicited advice from scholars in the field of Holocaust studies and attempted to integrate their sometimes divergent views into its work, taking into account established academic terminology as well as likely areas of interest to researchers. Standardized terms were sought whenever possible to ensure consistency with accepted terminology in relevant fields. However, Library of Congress subject headings were often too broad to service the sheer volume and depth of content to be indexed.

Therefore localized index terms were generated based on literary warrant. As the Foundation's catalogers began building the thesaurus they realized the need to further subdivide already existing subject areas. This was due to the extremely specific subject matter of the material the Foundation was indexing. As a result, certain areas of the thesaurus are much more detailed than others.

In the thesaurus there are two main types of relationships between terms: equivalence relationships and hierarchical relationships.

**equivalence relationships**

The thesaurus expresses equivalence relationships between synonyms and sometimes quasi-synonyms. This is achieved by designating one term to be the preferred term and assigning the other terms as non-preferred terms. This relationship is especially useful when dealing with terms that have more than one variant (particularly in the case of language variants). For example, if an end user wanted to search the archive for mentions of "the Joint" he or she would be guided to the preferred term.
Survivors of the Shoah

In the case of quasi-synonyms, the generic posting rule allows for equivalence relationships to be established between a class and its members. For example, in the thesaurus a Mohel, a Baal for equivalence relationships to be established between a class and its members. For example, in the thesaurus a Mohel, a Baal Tefilla and a Shaliah Tzibur (which are all specific types of Jewish ritual leaders) are designated as variants to the index term "ritual leaders, Jewish." Therefore by utilizing the process of generic posting catalogers are able to limit what otherwise would be an exponential growth of the vocabulary.

hierarchical relationships

In addition to the equivalence relationship the Foundation's thesaurus represents both the generic and whole/part hierarchical relationships. Using the generic relationship, controlled vocabulary is organized into user-friendly classes. For example the broadest classes are beliefs and perceptions, events, institutions and organizations, and places. Given the huge volume of material in its archives the flexibility of the generic relationship makes it easy for catalogers to decrease or increase the level of specificity in any area of the controlled vocabulary throughout the indexing process.

The whole/part relationship plays an equally important role in the thesaurus. High-level concepts such as "ghetto life", "religious life", and "world" are referred to as containers and house, narrower index terms that are part of the larger concept. For example, the "world" container, which contains all the geographic terms, is subdivided into continents; then countries; then states (where applicable) and then cities, ghettos, concentrations camps, etc.

Example 1

World

Asia

Europe

Poland

Cracow (Poland)

Auschwitz (Poland: Concentration camp)

Lódz (Poland: Ghetto)

North America

As more and more testimonies are indexed the biggest challenge is vocabulary control. A team of historians and historical authority supervisors discusses and defines each term added to the thesaurus. They ensure the historical validity of the terminology by using authoritative reference sources. When possible they import terms from standardized terminology such as Library of Congress Subject Headings or other authoritative reference works. Once a term is established it includes an authoritative source reference, a definition, and a scope note that defines its usage.

When terms cannot be verified in any authoritative source the historical team recommends whether to incorporate them into the thesaurus and, if they are added to the vocabulary, records the fact that no accepted reference work documents them. Terms considered "unverified" may eventually be verified by other sources or by the archive itself since the sheer volume of subject-specific information contained in the archive may evolve.
The indexing process

In order to grant end users access to portions of videotaped testimonies relating to specific matters of interest, each testimony is divided into segments (small narrative units) to which index terms can be assigned. Segments are usually about three to six minutes in length. They are selected based on the content delivered by the interviewee.

Segments are structured to have a logical beginning and end point — the break between segments occurs when the interviewee begins to describe a new experience. Typically segments are created to capture discussions of the interviewee’s family, his or her education, religious life, aspects of ghetto or camp life, etc.

The Foundation’s indexing methodology enables it to maintain consistency and control the level of specificity in indexing. While indexers exercise a certain amount of personal discretion in applying index terms based on the information provided in the testimony, controlled vocabulary and methodology guidelines help limit the level of subjectivity. In practice, indexers do not attempt to index every minor concept mentioned by the interviewee. Rather, indexers are trained to capture the major topics discussed. Therefore, the way in which interviewees discuss their experiences directly affects the assignment of index terms, e.g., a six-minute segment could have two index terms while a three-minute segment could have six (Figure 3).

In general, indexers apply two categories of index terms: geographical and experiential. Both categories can present certain challenges for the indexer. These challenges have the potential to increase subjectivity in assigning terms. The Foundation’s indexing methodology was constructed to counter this tendency. For example, the methodology instructs indexers to assign geographical index terms in two instances: when a specific location is being discussed or described in detail, and/or when events being described take place in a specific location.

Example 2

First segment: An interviewee describes her experience arriving in Auschwitz concentration camp. Second segment: The interviewee goes on to discuss a specific event that happened during the arrival process without explicitly stating the camp name.

In both segments, the indexer would use the term Auschwitz (Poland: Concentration Camp) as a geographic index term. The indexer would also assign experiential terms to reflect the experiences being discussed.

Example 3

An interviewee is describing his experience in Auschwitz while contrasting it to his life in the town of Oswiecim before the camp was established.

In this instance an indexer would choose the term Auschwitz (Poland: Concentration Camp) as the geographic term for the
Survivors of the Shoah

description of the experience in the camp and the term Oswiecim (Poland) for the description of his experiences in the town.

Not only do the Foundation's indexers create segments and assign index terms, but they also write an abstract for each segment that provides context for all of the assigned index terms. These abstracts are important as hundreds of segments can potentially have two or more of the same index terms; therefore it is necessary to enable end users to distinguish between segments and to narrow their searches. Indexers also compose a one-page testimony summary in order to provide context for the testimony as a whole. This enables end users to view a brief overview of the interviewee's experience. Without these two additional tasks it would be more difficult and time-consuming to sift through the large quantities of content contained in the Foundation's collection.

To ensure consistency, an indexed testimony goes through a two-tier review process. Indexers are assigned to teams headed by reviewers. These reviewers verify that index terms are assigned properly and ensure that the abstracts and summaries are generated with a consistent tone. Reviewers are assigned to Historical Authority Supervisors who not only oversee the review process but also help to develop and maintain the thesaurus.

Summary

The unprecedented scale of the Shoah Visual History Foundation's mandate has presented unique challenges in the development and implementation of the thesaurus and indexing methodology. Its purpose has been to enable end users at all educational levels to approach and explore the wealth of personal experiences recorded in the collection. It emphasizes to both its indexers and the end users that the index is intended to serve as a navigational device that leads to the interviewees' own descriptions of their experiences. The thesaurus and the metadata are not intended to be an end but only a means of arriving at the core resource: the actual digitized testimony.

The decision to index in such a detailed fashion has allowed the Foundation to be flexible and creative and has taken it beyond established indexing rules. As such the hope is to create an archive which will serve the public, including researchers and students alike.

The Shoah Visual History Foundation web site is at http://www.vhf.org. Mary Crystal (crystalm@pacbell.net), vice-president of ASI's Southern California Chapter, is a freelance indexer and journalist based in Pasadena, California.

MACREX INDEXING PROGRAM

Main office: MACREX INDEXING SERVICES, Beech House, Blaydon Burn, Tyne & Wear, NE21 6JR, England
phone and fax: 0191-414 2595 (international: +44 191 414 2595)
email: D.Calvert@macrex.cix.co.uk web site: http://www.macrex.cix.co.uk/

NEW FEATURES IN MACREX 7 FOR WINDOWS 95/NT (Features marked * are available in MACREX 7 for DOS)

- mouse control for
  - selecting entries on screen for editing - move directly from one entry to another
  - scrolling a line at a time on add and edit screens
  - selecting and changing menu options
  - saving and loading default files
- search and replace simplified *
- adding of entries in edit mode streamlined *
- pop-up menus *
  - scroll buffers for last 1000 entries and last 100 commands
  - explanations of error messages
- long file names accommodated
- direct access to Windows printer drivers
- default record length increased to 300 characters *
- maximum record length increased to 3000 characters *
- option to change record length while loading backup file
- screen length and width can be increased to accommodate more than 50 entries per screen

AGENCIES (for United Kingdom, Europe, Africa, South and Central America contact the main office above) for North America (United States and Canada): MACREX SUPPORT OFFICE, Wise Bytes, P.O. Box 3051, Daly City, CA 94015, USA
phone 650-756-0821; fax 650-757-1567; email: macrex@aol.com

for Australia, New Zealand and South-East Asia: MASTER INDEXING, 44 Rothesay Avenue, East Malvern, Vic. 3145, Australia
phone and fax (03) 9571 6341; email: mindexer@interconnect.com.au