Fortunoff Video Archive: New Initiatives for a New Digital Archive

[1]Kevin Glick and [2]Stephen Naron

Fortunoff Video Archive for Holocaust Testimonies, Yale University Library, New Haven, USA

1Kevin.glick@yale.edu
[2]Stephen.naron@yale.edu

Abstract. The Fortunoff Archive recently completed a number of important milestones including the digitization of its entire collection, the development of a digital access system, migration of legacy metadata, and the launch of a partner site program that provides remote access to testimonies at universities and research institutes. This paper will present an overview of these new digital initiatives.

Questo paper offre una panoramica su alcuni dei principali risultati raggiunti dal Fortunof Archive (Yale University) nel processo di digitalizzazione della sua collezione di videointerviste ai sopravvissuti della Shoah. Oltre alla digitalizzazione dei materiali, particolare attenzione viene data ai temi della migrazione dei metadati e dell'accesso remoto alle testimonianze conservate da istituti di ricerca terzi.

Introduction

The Fortunoff Video Archive is only now emerging from its analogue past. The Archive completed migration of its entire collection in December of 2015, after a 5-year in-house digitization effort using SAMMA solo technology[1]. It was an enormously complex and costly effort for an organization of our size, and the project's scale tested Yale University Library's capacities. Managing the data alone-more than half a PB-was a challenge. Nevertheless, this project helped push the Archive and the Library as a whole to planning the future of digital preservation, large scale storage, and access to AV special collections materials at Yale University Library.

[1] For technical specifications on SAMMA solo see https://docs.oracle.com/cd/E65703_01/en/solo_admin_4.2/solo_admin_4.2.pdf

Brief Collection History

For those unfamiliar with the Fortunoff Video Archive, the collection's roots stretch back to 1979 when a small group of survivors, children of survivors, academics, and a local television personality joined forces to create the Holocaust Survivors Film Project in New Haven, Connecticut. It was the start of the first sustained effort to record the testimonies of survivors, witnesses, and bystanders on video- tape. Sustained means that the Archive is still recording survivor testimonies at Yale's studio in New Haven. The original 183 recordings were deposited at Yale University Library in 1981, and the Archive opened its doors a year later. Over the years, the Archive expanded its recording effort with the help of more than 35 affiliate projects worldwide. These affiliates signed an agreement with the Archive, which then trained the affiliate staff according to the Archive's specific interview methodology. Each affiliate then sent a copy of their master tapes to New Haven, where the Fortunoff Archive staff cataloged and cared for them. As part of this affiliate agreement, the Archive and the affiliate generally share copyright, so the resulting recordings belong to both Yale and the affiliates. The collection currently consists of around 4,500 testimonies in about a dozen different languages, encompassing more than 12,000 hours of recorded material.

An Incomplete Digital Transition

The most significant change this digital turn has brought to the Fortunoff Archive is the launch of our new online access system. Until March of 2016, researchers were still using VHS tapes in the Yale University Library's Manuscripts and Archives reading room. The online access system has allowed the Fortunoff Archive to break down some of the physical barriers that have restricted use of the collection in the past. To initiate access to users outside of New Haven, the Archive launched our "partner site program." Former affiliate projects, institutions of higher learning, libraries, archives, or Holocaust research centers can request to become a Fortunoff partner site and gain remote access to the entire collection free of charge. Partner sites sign a Memorandum of Agreement with the Fortunoff Archive, and then provide the Archive with an IP address. Researchers can then search the Archive's online catalog, register in Yale's special collections request system (Aeon), request a testimony, and once approved by Archive staff, view the testimony at the approved partner site. We now have almost a dozen partner sites active worldwide, and more than 98 unique users from several different countries who have requested around 450 different testimonies since the online access system was launched. However, we are still very much in midtransition from analogue to digital. We can categorize the challenges and opportunities the Archive faces during this transition into two main areas: Challenge 1: Metadata and Discovery , and Challenge 2: Access and Systems Integration .

Challenge 1: Metadata and Discovery

Over the years, the Fortunoff Archive has created a wealth of descriptive, technical, preservation, and administrative metadata for every testimony. Each interview has a 200+ word summary, and dozens of Library of Congress descriptive subject headings[2] representing the geographic, agent, and topical content of the testimony. While detailed, the Archive's descriptive metadata is rooted in 20th century American archival and library practices. In addition, it does not necessarily share a taxonomy or cataloging rules with other testimony archives. The Fortunoff Archive uses MARC-21 and the controlled vocabulary of Library of Congress subject headings, but other institutions holding testimony have other practices. Furthermore, the metadata about the collection, technical, administrative and descriptive, is spread across multiple systems. Kevin Glick, who is responsible for much of system development at the Fortunoff Archive, led a project, with the help of an external consultant, to map this data and consolidate it into two systems: 1) ArchivesSpace, the archival description and collection management database for archival materials utilized at Yale; and 2) Preservica, our digital preservation system. Our hope is that in the end, this metadata consolidation will open up new paths towards sharing and reusing our collection metadata. Even before this migration the Fortunoff Archive began experimenting with shared metadata, in particular with the help of one of its first partner sites, USHMM. Not only is the Archive providing access to the en- tire collection on site at USHMM, but we are working with staff at USHMM, under the leadership of Michael Levy, to import Fortunoff Archive metadata into USHMM's discovery system, Collections Search[3]. USHMM periodically harvests Fortunoff catalog record records and ingests them into collections search as a ”Special Collection.” By including the Archive's catalog records in this search, USHMM's portal allows researchers to search across collections that have never been collocated before. For example, for the first time, you can conduct a search for interviews of survivors of the Lodz ghetto who gave testimony to both USC Shoah Foundation and Fortunoff. There are still a number of challenges we need to address with this metadata sharing project. The Fortunoff Archive still has not determined the best way to dynamically provide updated catalog records to USHMM. Our hope is that the Archive's move to ArchivesSpace, which has more advanced features like an API, and an export function for bibliographic and EAD records, will help us find a solution. Another challenge to this type of data sharing initiative is the longstanding archive policy to truncate last names of survivors in all public-facing materials. Until our stakeholders agree that it is acceptable to release last names of survivors, and/or find some other creative ways to ”connect” testimonies from the same survivor in different collections (a central shared unique identification system for example), researchers will still find it impossible to do a global search for all the different interviews one individual may have given. For example, if you search the USHMM collections for Simon Srebnik, you won't find Yale's catalog record for his interview. Until we find a way to rectify this, there will be many missed research opportunities. Discussions are underway regarding the use of survivor's last names, and we are optimistic that some compromise will be possible to help us enable the type of data sharing that will make it simpler for families and

[2] Library of Congress Subject Headings, http://id.loc.gov/authorities/subjects.html

[3] https://collections.ushmm.org/search/

researchers alike to locate testimonies. While the success of sharing metadata with USHMM to enhance discovery of testimonies is certainly something to celebrate, it still underscores certain limits on the way we have envisioned how researchers search in our collection. Our discovery system is based primarily on the American library OPAC model that was designed for books. It is a flat, simplistic and highly mediated form of discovery. You can search across catalog record data from different testimonies, but you cannot search across any accompanying material like interview summary notes, transcripts (if they exist), or geographic coordinates. And there are no truly interactive links between testimonies, either based on common interview participants, descriptive elements, or content- beyond shared LC subject headings. In addition to the testimony-level catalog records, each testimony is accompanied by a set of summary notes. Essentially, these notes are a very summarized transcript of the testimony, in the first person, in English, regardless of the original language, with timestamps embedded in the text every 5-10 minutes. They come in a variety of shapes and sizes, formats, and quality. These notes have been produced by student workers over decades-at times, applying less than rigorous quality standards. These notes were used by archivists to catalog the testimonies, but are also used by researchers as maps to navigate testimony content. To prepare these notes for use in a discovery or access system, we needed to create a single, uniform, structured format. Another complication is that these notes were created using the visible timecode on the Archive's VHS use copies - not the the inherent time-code from the original master recordings. This means that the timestamps in the notes do not match the timecode of the recently digitized master tapes. The discrepancy varies from tape to tape, so there is no way to synchronize the notes with the master video algorithmically. To remedy this, the Archive launched a project using an open source product developed at the University of Kentucky for its Oral History program. The system is called OHMS, which stands for oral history metadata synchronizer, and the idea was to achieve two goals simultaneously: to reformat all the notes into a single, structured XML schema, and to correct all incorrect time-codes so that the text points accurately to the appropriate segment of the digital video. To do this, students use OHMS to sync the notes with the digital video. They do not need to watch the entire testimony, but rather ”skim” through the testimony to locate the points in the video that corresponds with the start of each paragraph in the notes. They can then tag the video where those paragraphs begin, thereby synchronizing the text to the moving image. Since we started the project last year we have OHMSed over 3,500 tapes. We estimate the project will take at least another two years to complete at this pace. The expected benefits of having this cleaner, structured data is that it will be much easier to include these notes in some future index that will allow researchers to discover testimonies by searching across these notes as well as the testimony-level descriptive metadata But our ambitions to enhance discovery, and the ways in which researchers can interact with the collection goes beyond standardizing legacy notes. The Archive intends to create and integrate verbatim transcripts for the entire collection into our discovery and access systems. In fact, the Archive recently received a grant from the Krieble Delmas Foundation to transcribe the first 183 testimonies recorded by the project. This project, while small in scale, will help us develop workflows and standards for producing verbatim transcripts for larger parts of the collection. Not only will these transcripts eventually be integrated into discovery and access systems, but they will also serve as the base documents for a number of other new initiatives: 1) the launch of a dynamic online critical edition series of annotated testimonies, created by our first Hartman postdoctoral associate, Sarah Garibova, and 2) a testimony transcript analysis tool to be developed by the Archive's first Digital Humanities postdoc, Gabor Toth. The critical edition series is essentially a dynamic presentation of testimonies, with interactive transcripts, accompanied by annotations created by the postdoctoral associate to help elucidate the testimony's narrative. The scholar will provide detailed explanations and context, thereby opening up the testimony to a broader readership. This series will be openly available, hosted on- line, and allow viewers to toggle back and forth between the transcript, annotations, and the underlying video. Gabor's transcript analysis tool will help us explore a large sample of transcribed interviews from different collections. As a whole, we hope that Gabor will be able to work with a sample of 3000 interviews from the USHMM, USC Shoah Foundation, and Fortunoff collections. The tool will have different functions facilitating a data driven exploration of the interviews, and the underlying memory. First, there will be a simple transcript reader, where users can browse and read the transcripts, and if possible listen to the original interviews. The tool will also help readers to find connections between different testimonies. In practice, this means that users will be able to click on any word in the transcripts, and find its occurrences, or the occurrences of its synonyms and antonyms, in other testimonies. The exploration of the interviews will be also supported by different machine generated indices. We will for instance extract names, place names, objects, and key topics from the transcripts, and let readers use them to browse the interviews, and again to uncover hidden connections. It is still an open question, where and how exactly we will make this tool, and the critical edition series, available to the public, and how they will tie into our current discovery and access systems.

Challenge 2: Access and Systems Integration

This brings us to the Archive's last set of challenges: access and systems integration. Although an enormous leap forward, our current access system is a work in progress, leaving something to be desired in terms of functionality. For instance, as currently configured it requires users to request access to each testimony separately, and one cannot easily move between testimonies. Also, each testimony request must be approved manually by Archive staff, which slows down the request process, and makes it difficult to ensure researchers at locations will have a chance to view everything they want to view when they want to view it.

There is also a significant amount of associated materials in the Archive that are not yet integrated into the access system, such as pre-interview questionnaires, and other related ephemera and records. Finally, the OHMS project, Delmas grant, the critical edition series, and the transcript analysis tool (discussed above) all raise questions about how researchers will search, navigate and consume all of this rich content, both ”inside” the testimonies and across multiple testimonies. How will they move between different views, tools, and systems? The Archive's systems are already highly complex, and largely disconnected with connections made manually or through minimal automation. There is very little direct dynamic systems integration between the Archive's management, preservation, access, and discovery systems.

The following briefly described the the current access system to give a sense of its complexity: Discovery occurs via the library's online public access catalog, Orbis[4] where researchers search across testimony catalog records to locate materials of interest. Each catalog record contains an Open URL link that directs a request for a single testimony to another system, Yale's Special Collections re- quest management system, Aeon. The user is then required to register to complete the request. Once the request is submitted, and approved by the archivist, the user receives an email informing them to report to a partner site to view the testimony. Access to the system is limited to specific IPs or IP ranges at approved sites. Once the researcher is at a site, on a computer with an approved IP, she logs into her Yale Special Collections Aeon account, and clicks on a link in the testimony request. This link directs her to a Drupal webpage (if you have been counting, you will notice that users have now been pushed into at least the third or fourth system, all of which are separate from one another, and have a distinct look and feel). The Drupal page, then dynamically pulls together the streaming video of the testimony, which is hosted on Kaltura's platform, as well as the time-coded summary notes from a web-service so that the notes dynamically display while the video is playing.

Conclusion

The mission ahead is to transform this labyrinth of Yale enterprise, vendor hosted, and custom tailored systems into a seamless experience for the researcher, as well as integrating the new digital initiatives from our postdoctoral fellows. All of this while aiming to build ”something” that plays nicely with others, allows for metadata sharing, and opens opportunities for cross-institutional cooperation. In the weeks and months ahead, the Fortunoff Archive intends to make significant progress, both internally and in collaboration with others, to streamline the process whereby re- searchers can discover and access materials within our collection, and in conjunction with other similar, aggregated collections.

Last URLs access: June 21, 2017

[4] A listing of all unrestricted catalog records of testimonies can be discovered with this string, https://goo.gl/vH9JY4 accessed June 21, 2017.