Umanistica Digitale - ISSN:2532-8816 - n.4, 2019
M. Levy – Some Perspectives on the Practice of Sharing Collection Data
DOI: http://doi.org/10.6092/issn.2532-8816/9039

Some Perspectives on the Practice of Sharing Collection Data

Michael Levy

United States Holocaust Memorial Museum, Washington, DC, USA
mlevy@ushmm.org

Abstract. This paper explores diverse areas of considerations regarding sharing data relating to Holocaust collections, from institutional, motivational, ethical, and technical points of view.

Questo paper affronta il tema della condivisione dei dati relativi alle collezioni di documenti sulla Shoah, dal punto di vista istituzionale, motivazionale, etico, tecnico.

With Whom Should We Be Sharing Data?

Collecting institutions describe their holdings and gather and codify information that they hope will be of use for researchers. Holocaust institutions maintain collections that serve a very wide variety of purposes. People accessing Holocaust collections may include students of all ages from elementary, secondary, college, graduate and postgraduate levels; scholars, authors, documentarians, and journalists; those doing family and other genealogical research; those seeking evidence for various legal needs; those informing themselves on a broad variety of areas; and--more and more in the last few years--data scientists and digital humanities researchers. We should assume that over time, people will come up with research questions, tools, and techniques not yet imagined.

The range of benefits to society and to institutions accrued through data sharing activities may be difficult to measure, although each of us involved in our institutions will surely have many anecdotes. Most institutions holding Holocaust-related collections have supported scholarly research and publishing and are well aware of resulting scholarly output. In addition, those of us who have experienced firsthand the overwhelming emotion experienced by a survivor or family member when viewing documents relating to themselves or to a close family member for the first time, or after decades, will never forget the power of documents embodied in our collections.

In this paper I will focus mainly on bulk sharing of collections metadata. Holocaust collecting institutions also engage in the extremely valuable practice of copying archival and media

collections between institutions, and between institutions and researchers. In addition, the USHMM has been the beneficiary of many indexed names acquired through many channels that now make up the Holocaust Survivors and Victims database, which provides the ability to search approximately 34,000 cataloged lists and 7 million name records. Here I focus mainly on collections metadata other than the name lists and indexed name-related data.

Collections catalog data is most often human-generated descriptions which are based on cataloging standards and practices, and include such data types as dates, summaries, descriptions, subject headings, formats, genres, and other authority-based cataloging terms, indexed names, dates, and other data transcribed directly from all kinds of materials, transcribed language from spoken media or text, and administrative and organizational metadata.

After having generated and collected these many types of data about our collections, institutions ask, "With whom should we be sharing data, and for which purposes?" I quote below from the Society of American Archivists Core Values Statement and Code of Ethics:

Archivists promote and provide the widest possible accessibility of materials, consistent with any mandatory access restrictions, such as public statute, donor contract, business/institutional privacy, or personal privacy. Although access may be limited in some instances, archivists seek to promote open access and use when possible.

In short, the default position is to share: i.e., share, unless you are compelled not to share by a restriction with which the institution must comply.

When Sharing is not the Default

Issues relating to privacy and dignity and the control and protection of information about persons living and dead, in particular with records relating to the Holocaust vary widely between countries and institutions. The differences between US and European attitudes, norms, practices, and regulations is well known. EU regulations regarding data relating to persons are in general more stringent than those in the United States. Happily, the EU General Data Protection Regulation (GDPR) includes a specific reference to the Holocaust aimed at supporting access to such records.

National laws enforce copyright restrictions for published materials. Sometimes institutions collect digital materials acquired or copied from other institutions under agreements that greatly restrict the rights of the receiving institution with respect to sharing the data. Some records that, for example, relate to certain experiences that people have been subjected to or that relate to behaviors under extreme circumstances may be considered an affront to human dignity, and these records are often restricted in an attempt to honor the affected persons and their descendants. And yet one more reason to restrict data sharing can occur when people currently living in political environments in which evidence of their ancestors having been identified as being a member of a group targeted for persecution may, in the current day, cause them to fear repercussions, so there could be a personal safety consideration.

In addition to the barriers listed previously, embedded cultural forces within collection-holding institutions tend to work against sharing data. Below I list some cultural forces that power reluctance to share data. The points below are not direct quotes but are my own impressions of arguments that may be behind some impulses against sharing data.

- "We've never done this before and cannot predict exactly what might happen if we were to share data. Status quo is safer."
- "We worked hard to collect, collate, describe this data; we can't just give it away. This data is an institutional asset, a competitive advantage that belongs to us and helps differentiate us from other similar institutions."
- "If our digital collection is available online, researchers will no longer need to visit our archive in person, and could affect our status or level of support."
- "Making our data easily accessible would expose quality issues, exposing inconsistencies or even inaccuracies to others. This would require us to respond to errors and suggestions for correction, and we do not have the staff, support, and technology to do that."
- "This data embodies the evidence of such deep significance, such profundity, and represents such deep spiritual meaning, that we must not simply place the related textual data into the hands of anyone. We know that there are those who wish to minimize or distort Holocaust history, and we need to do what we can to prevent possible misuse of records and disrespect of memory."
- "We need to maintain control over the interpretive narrative, to tell the story in ways that we feel honors the gravity of this unique historical event."

When *closed* is not mandated, *openness* promotes broadest use and reuse, and is, in my opinion, generally worth the risks. In the balance of open versus closed, open allows for broadest reuse and collaboration.

Some Methods of Data Sharing at the USHMM

Any online public access catalog is intrinsically one means of sharing data from an institution's collection.

The United States Holocaust Memorial Museum's Collections Search is such an example. The USHMM developed a highly customized version of an existing set of open source systems including the catalog search and discovery system called Blacklight. The history of the collection activities at USHMM varied tremendously over the years. Different parts of the collection were gathered by different people at different times for different purposes and for different audiences. Systems used to create and maintain the catalogs also varied widely. For example, the Library used commercial integrated library management systems, which enforced standard MARC-based data representation. The unit at the Museum who maintained the

photographic reference collection and associated services have used an entirely bespoke system, using bespoke keywords appropriate for their intended audience. The unit working with historical film also, until very recently used another bespoke database with bespoke descriptive elements. The work of cataloging three-dimensional historical objects is done within a commercial collections management system that also provides a cataloging subsystem. Oral histories and archival collections use the same system, but configured to use different fields in somewhat different modes. Keyword systems comport with standard cataloging taxonomies but the system is not entirely rigid in ensuring absolute conformance with those standards.

The initial approach to the creation of archival finding aids was in a time of paper and typewriters. The advent of word processing software allowed computerization. At our institution, the ratio of archival finding aids to staff, both trained archivists and technicians, and institutional priorities, has not as yet allowed us to convert archival finding aids to EAD, which is the leading XML-based standard method of creating, editing, and sharing archival descriptions. Nevertheless we found ways to make the content of the finding aids, as well as of oral history transcripts and other textual resources available to users through indexing that unstructured, or "barely-structured" text (finding aids, transcripts, etc) along with the more structured catalog records using a search engine (Apache Solr) and making that available.

Before 2013, the USHMM's public website contained Collections-related materials primarily in three separate areas of the Museum website. The USHMM Library cataloged publications data using a commercial integrated library management system, which provided a user friendly public access catalog. A subset of Museum items and some oral history descriptions were also included into the Library's public catalog interface. In two other, disparate sections of the Museum website, the Photo Archives and the Film and Video Archive provided access to digitized photographs and descriptions and to historical film segments.

- The decision to provide a single user interface to all types, formats, genres, and forms of material was based on a conjecture that doing so would provide the most overall benefit to the widest variety of users. Collections Search presents a single search interface, behind which are numerous distinct metadata sources, textual files, databases. All of these are integrated nightly into a Solr index for search and delivery.

The major components include:

- MARC based records from the Museum's Library cataloging system, which are indexed into the Solr search engine nightly using SolrMarc, an open source project
- Metadata from the Museum's collections management system's cataloging module (comprising descriptions of Document (Archival) records; Oral History records; and Object (including Art) records; and historical Film. Each of these sources has its own set of metadata fields.

- A bespoke database system, based on a free open source PHP/MySQL system, SuiteCRM, links cataloged items with media files, and these data and links are also indexed along with the other elements for searching, faceting, and display.
- Archival finding aids; Oral History transcripts and notes; and Film transcripts and other textual materials mostly in Microsoft Word, are converted to PDF, the text is extracted and converted to XML, and those indexed records are merged with their respective catalog records so that they are searched and delivered together.
- Photograph cataloging is performed with a standalone bespoke desktop database, whose records are transformed into XML and indexed into Solr.
- As the Museum's paper archival holdings are being digitized, they are ingested into an open source DAMS system ResourceSpace. This is now being made available using IIIF interfaces and using the Universal Viewer (more details below).
- Oral history collections metadata from two collections holding institutions have also been made available through Collections Search (more details below)

The timing of this work turned out to be very fortuitous for Museum participation in the EHRI portal. Because the metadata had already been packaged and integrated for the purpose of building the online catalog Collections Search, the level of effort required to share the metadata with EHRI partners was essentially zero, and became an excellent example of unanticipated data reuse.

Benefits of Sharing Data

Much of the data shared by USHMM does not comport to international standards. However, the Museum decided to share this data with EHRI in the form it was in at the time. The line of reasoning is as follows: this data is useful and is searchable using commonly available searching mechanisms, such as Solr. It would be impractical to wait until all data was normalized according to nationally- or internationally-recognized standards. The benefits of making the data available as soon as possible are twofold: 1) those exploring our collection could access the materials sooner, and 2) researchers who have tried to reuse our data have shown us where our data was confusing, inconsistent, and incomplete, and this has been beneficial to us, as it prompts us to better document or to clean up our data.

Rufus Pollock, founder of Open Knowledge International, wrote "The best thing to do with your data will be thought of by someone else" and "The set of useful things one can do with a given informational resource is always larger than can be done (or even thought of) by one individual or group." This is another pithy way of suggesting the power of openly sharing data with other individuals or organization who may possibly develop useful approaches, research questions, and insights beyond those that may have been developed by your own organization and its members.

Sharing bulk data has been practiced in libraries for decades, at least since the 1960s. Libraries pioneered collaborative cataloging practices because many libraries owned the same books, and the economic and practical disadvantage of each library cataloging the same books was obvious. Libraries have a long history of developing highly standardized methods of cataloging, using standardized, uniform formats such as MARC. The concepts of collaborative and cooperative cataloging practices for in cultural historical museums is a much less well established practice. Items in museums and archival collection tend to be one-of-a-kind, and different institutions tend to approach cataloging methodologies somewhat differently.

As an institution shares its collections-related data more widely, the institution learns about what is valuable to others. This then can feed back to the institution the information that the data is valuable to others. I am reminded of an old saying among real estate agents. Once a potential buyer starts to complain about features of the house (e.g. the color of the bathroom fixtures is all wrong, the carpet is ugly) then the agent knows they are good prospects and a sale is more likely than if they remain silent. The criticism shows interest in the material; the opposite is apathy or indifference. At USHMM as we have begun to share bulk data, for example with EHRI partners, we have become more aware of issues relating to data quality, and over time we may be able to incorporate this feedback into our cataloging practices. As the institution becomes more aware of the potential increased benefits of metadata reuse, perhaps more resources may be provided for data cleaning and transformation into standardized formats. As noted previously, the Museum shared bulk metadata with EHRI for use in the portal, with very positive results. It was, however, very apparent that the bespoke, nonstandard nature of the metadata formatting caused may difficulties for those doing the integration work.

In addition to sharing metadata, the Museum has made great strides in sharing media digitally, including time-based and image media. The USHMM has cataloged and made 25,000 hours of oral history audio and video, and about 2,000 hours of historical film available on the web. In addition, some percentage of the media is available only onsite on the Museum premises.

In addition to sharing our own catalog and transcription metadata with other institutions and individuals, the Museum has been the beneficiary of metadata provided by other institutions. In approximately 2011, the USC Shoah Foundation's Institute for Visual History and Education shared spreadsheets containing basic biographical information regarding approximately 52,000 interviewees with the USHMM for the purpose of integrating that data with our catalog, and since then and at the time of this writing, a web search for an interviewee's names very often results in USHMM Collections Search. In late 2016 representatives of the Yale Fortunoff Video Archive for Holocaust Testimonies shared a bulk export of MARC-based catalog data of their 4,500 interviews with USHMM with an intention that those records be made searchable along with the rest of the USHMM catalog. At the same time, the Fortunoff interviews are being made available to researchers onsite at the Museum premises. Based on the earlier work required to integrate metadata from various sources into a single search an access user interface, it was relatively easy to integrate the Shoah Foundation

interviewee biographical data and the Fortunoff data into Collections Search.

A Better Way to Present Hierarchically-Described Archival Collections

In April, 2017, the USHMM Collection Search launched an improved and more sophisticated method of providing access to hierarchically-described archival collections materials, and to date has presented over 700 collections of digitized paper archives on the web, comprising some 270,000 page images. The new method uses IIIF (International Image Interoperability Framework) interfaces to provide access. Archivists arrange many collections in hierarchical levels: series, sometimes subseries, and file level. The IIIF concepts and support for hierarchically organized Collection entities, from "Collection" to "Manifest" to "Sequence" and to "Canvas." These levels correlate roughly to collection, series, or subseries; file or folder; and item, image, or page. The USHMM began using ResourceSpace, a free, open source PHP/MySQL digital asset management (DAMS) system to manage these digital assets according the arrangement determined by the archivists, who process, arrange and describe the collection. The USHMM team developed a set of IIIF interfaces to the MySQL database behind ResourceSpace, along with other configuration and data processing steps to allow web presentation of the IIIF collections and interfaces. This work was supported by EHRI-2.

USHMM selected one of the open source IIIF viewers because it provides an excellent visual user interface combining the hierarchical nature of archival collections. The interface selected is the Universal Viewer (UV). The UV development had its origins in work done at the Wellcome Library and was later generalized as the Universal Viewer.

USHMM hopes that not only will the implementation of these IIIF-based presentations of archival collections provide a useful tool for remote users to view and browse through archival collections. We hope also that this can provide a platform for other innovative reuse of these digitized paper collections. Use of the standard IIIF interface would allow other web tools to present USHMM-server-hosted images and metadata. This could serve as a very convenient and flexible method to support innovative use and reuse of our collections by external parties.

For example, IIIF could provide a framework for crowdsourcing projects. Because this is one area in which interfaces are widely shared in a growing open source development community, new tools may be able to be integrated fairly easily. The subset of USHMM archival collections materials that are currently provided through IIIF interfaces could, for example, be crowd transcribed. To complete the circle, the transcribed text could then be used to further enhance access to the archival material through an enriched search and display interface.

The USHMM intends to continue to explore new and innovative ways of making our data and metadata as useful as possible through as many means as possible, while respecting ethical and legal principles and agreements, in order to foster and encourage scholarship and study of the Holocaust. This work will never be finished, and one may expect that innovative data sharing practices will only accelerate as the activities become more accepted and the benefits become

more easily appreciated and more easily achieved.

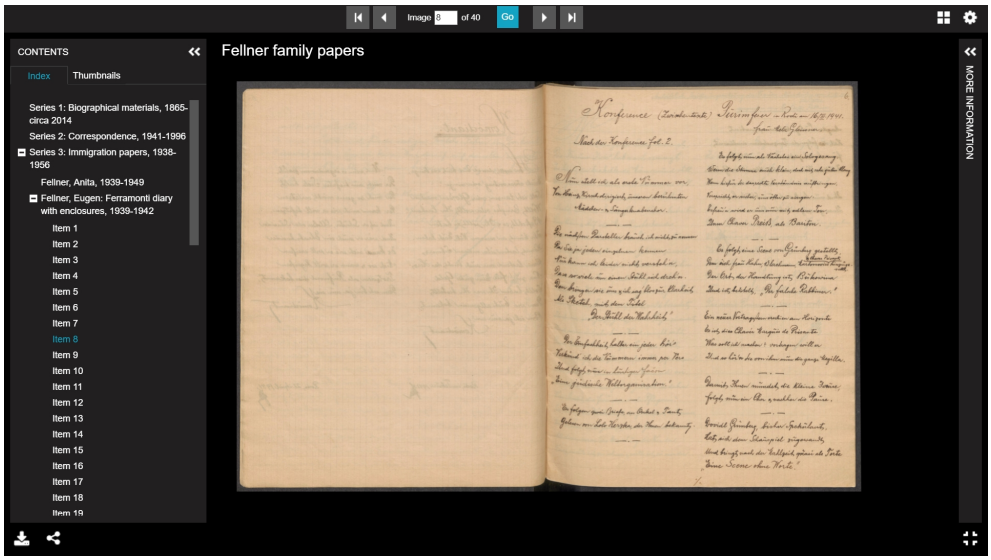

Figure 1: A page from a USHMM Archival Collection,IIIF/Universal Viewer

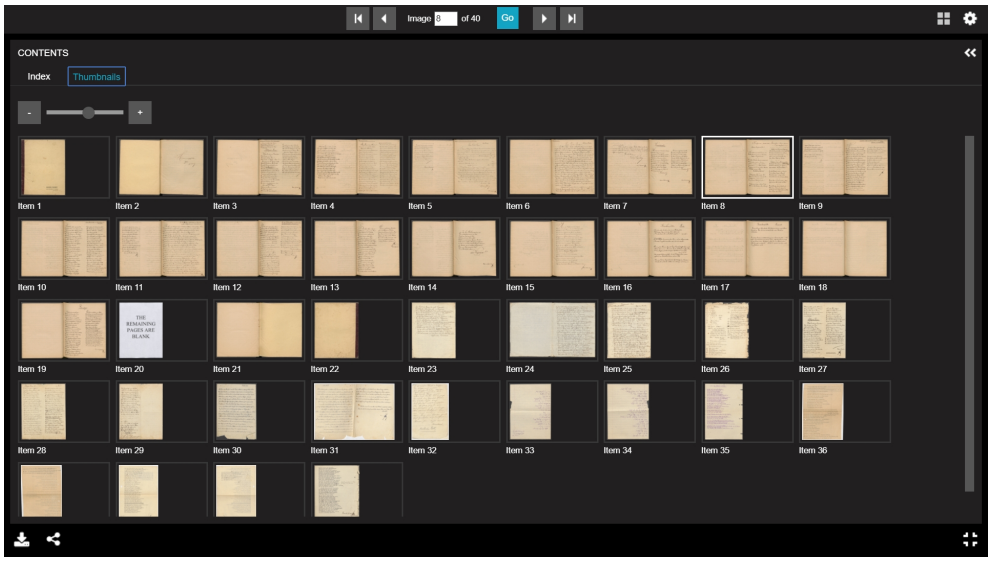

Figure 2: USHMM Archival Collection, IIIF/Universal Viewer, thumbnail view

A Way to Provide Harvestable Data

The Blacklight software provides two methods of harvesting metadata from any detail page: one is through content negotiation, which is part of the HTTP protocol. If a user agent requests JSON through adding “Content-Type: application/json” to the HTTP request header, the Blacklight software will provide all of the pertinent metadata relating to the page being viewed, formatted as JSON. A second method that is simpler to anyone to demonstrate is to simply append ".json" to the end of any URL page. For example, the Fellner Family papers collection, Accession Number 2015.563.1, the URL is: https://collections.ushmm.org/search/catalog/irn531331

Any web user may simply append ".json" to the URL, i.e. may request: https://collections.ushmm.org/search/catalog/irn531331.json

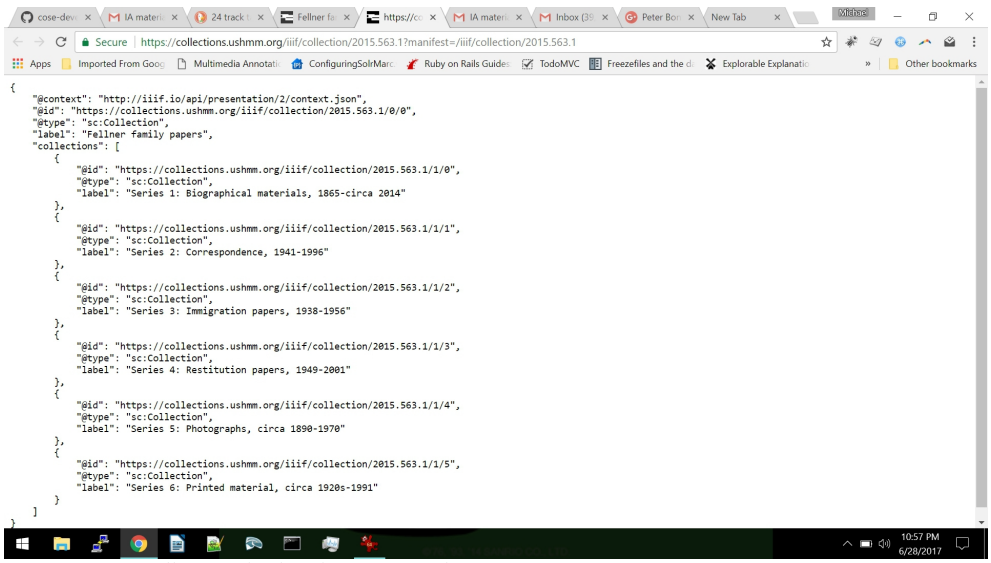

Figure 3: IIIF Collection displayed as JSON in browser

This will result in all of the metadata relating to that item being returned to the browser, which can then be studied by a user or could be harvested automatically through a harvesting script. Another example with an oral history interview: the oral history record and video can be accessed here: https://collections.ushmm.org/search/catalog/irn517852

The textual data comprising the record along with the full text of the interview is accessible here in JSON, a machine-readable format: https://collections.ushmm.org/search/catalog/irn517852.json

This includes text extracted from the Word or PDF finding aids and other metadata.

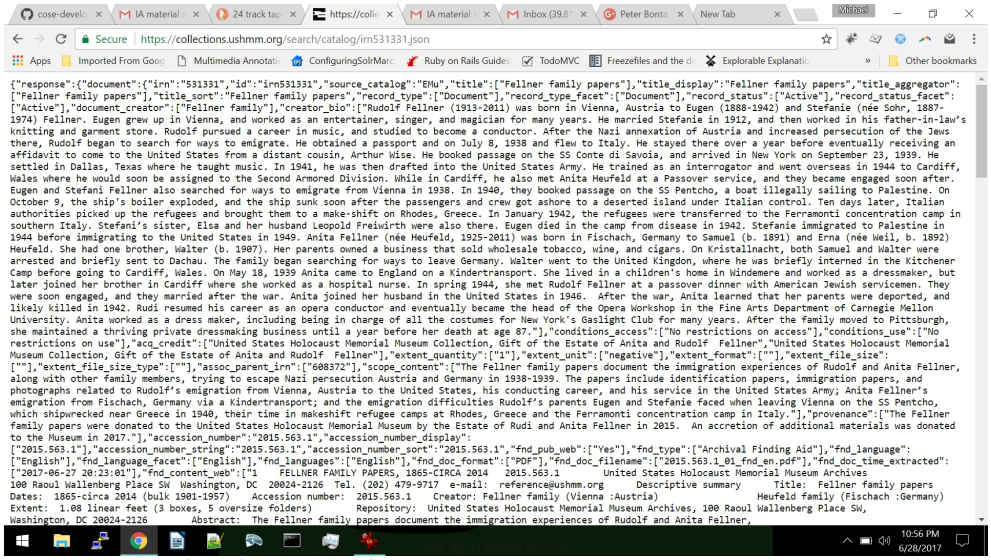

Figure 4: Display of text from archival finding aid

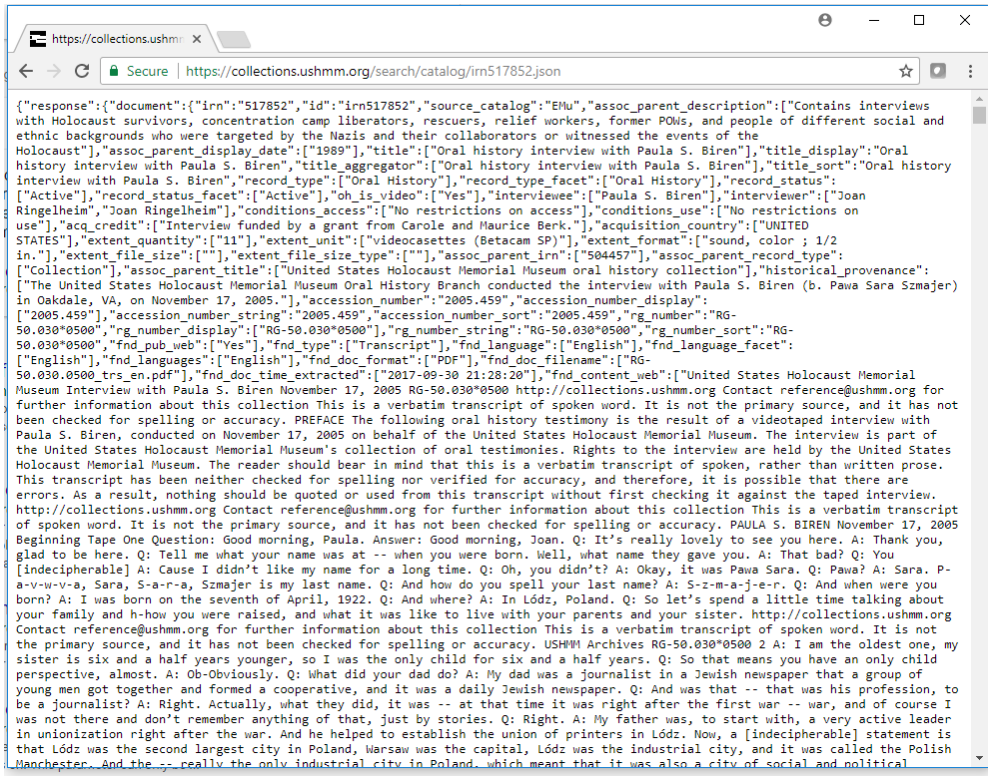

Figure 5: Display of JSON displaying text from oral history testimony transcript

In the near future USHMM may investigate better methods of advertising the availability of this metadata for use by digital humanities or other researchers.

Other EHRI projects support development of software and outreach to collections-holding institutions that may be interested in publishing their own collections metadata and/or submitting their collection metadata to the EHRI search portal. The USHMM does not currently publish OAI-PMH nor Resourcesync interfaces to their metadata. Through USHMM's participation with EHRI, there has come greater motivation to participate in this type of process, and USHMM has begun to engage with EHRI software for publishing catalog data through ResourceSync protocol. There is reason to expect that over time USHMM may publish data through this method. In addition, within USHMM there has been a good deal of discussion regarding making metadata available for wide research uses through posting bulk metadata on some well-known site. Although there is no concrete plan at this time, there is also reason to expect that this could be accomplished soon.

References

1. "SAA Core Values Statement and Code of Ethics. Core Values of Archivists", http://archivists.org/statements/saa-core-values-statement-and-code-of-ethics
2. "EU General Data Protection Regulation (GDPR)", http://www.eugdpr.org/
3. "Reference to Holocaust in GDPR" reported in International Holocaust Remembrance Alliance, https://www.holocaustremembrance.com/media-room/stories/brussels-includes-reference-holocaust-gdpr
4. "Brussels Includes Reference to the Holocaust in the General Data Protection Regulation," https://ehri-project.eu/brussels-includes-reference-holocaust-general-data-protection-regulation
5. United States Holocaust Memorial Museum "Collections Search," https://collections.ushmm.org/search/
6. Blacklight Project, open source Ruby on Rails software, http://projectblacklight.org/
7. MARC standards, https://www.loc.gov/marc/
8. EAD, https://www.loc.gov/ead/
9. Apache Solr, http://lucene.apache.org/solr/
10. SuiteCRM, open source CRM, https://suitecrm.com/
11. ResourceSpace, free open source PHP/MySQL software, https://www.resourcespace.com/
12. International Image Interoperability Framework™ (IIIF), http://iiif.io/

13. Universal Viewer, free open source IIIF-based viewer, http://universalviewer.io/
14. USC Shoah Foundation, The Institute for Visual History and Education, https://sfi.usc.edu/
15. Fortunoff Video Archive for Holocaust Testimonies, Yale University Library, http://web.library.yale.edu/testimonies

Last URLs access: August 30, 2017