Umanistica Digitale - ISSN:2532-8816 - n.4, 2019
O. Tolokonsky – Integration of Heterogeneous Shared Data: Yad Vashem's Perspective as a Data Aggregator
DOI: http://doi.org/10.6092/issn.2532-8816/9041

Integration of Heterogeneous Shared Data: Yad Vashem's Perspective as a Data Aggregator

Olga Tolokonsky

Yad Vashem, Archives Division, Jerusalem, Israel
olga.tolokonsky@yadvashem.org.il

Abstract. This paper explains Yad Vashem's role as a data aggregator, describing its experience in adapting to the changes in the ways data are shared and procured in the present. The challenges of integration of heterogeneous shared data and the workflows and processes devised in Yad Vashem Archive in order to deal with this task are presented. These processes are then illustrated by discussing a data integration use case.

Questo contributo illustra le attività collegate all' aggregatore di dati Yad Vashem di Gerusalemme. In particolare descrive l'esperienza di adeguamento all'evoluzione delle modalità di raccolta e condivisione dei dati, l'integrazione di dati eterogenei condivisi, il workflow e i processi messi a punto fino ad oggi. I temi sono presentati illustrati alla luce di uno specifico use case.

Yad Vashem – Data Aggregator

Yad Vashem, The World Holocaust Remembrance Center was established by the Israeli Parliament. The Yad Vashem Law, which was enacted in 1953, determined that among its other missions, the task of Yad Vashem is "to collect, examine and publish testimony of the disaster and the heroism it called forth…". Right from the start it has begun collecting documentation arriving from various sources. These include state archives and public institutions as well as memorial initiatives and research endeavors.

One of the central aims of the Archive is to preserve the documents for future generations, and to allow an easy and comprehensive access to the assortment of collections in its care. By doing so the Archive facilitates open research, commemoration undertakings and educational programs.

Over the years, and especially since the fall of the "Iron Curtain", the Archive has been collecting and copying Holocaust related documents that have been kept in various archives in

Europe and throughout the world.

Currently, Yad Vashem houses a major collection of Holocaust related documents. What sets this collection apart from others is the fact that its materials span a vide period of time: not only the years of persecution, but also materials about Jewish life before the Holocaust and its effects until present time. Another reason for its uniqueness is its multilingualism. The persecution of Jews during the Holocaust being perpetrated across most parts of Europe brought about the accumulation of materials in various European languages.

Cataloguing

Cataloguing system

In order to achieve its aims and goals to make its collection accessible to the public either through an on campus portal or over the Internet, the Archive needed to catalogue its materials in a comprehensive manner. To that end, nearly twenty years ago, an "out of the box" SQL relational database management system was selected. [1] This system allows centralized management of all the information in the organization thus enabling the Archive to catalogue its vast amounts of documentation in a uniform and consistent manner. The software requires strict adherence to predefined protocols and procedures. It forces the Archive to produce rigidly structured records. This is necessary because most of the cataloguing is done by using controlled indexing language, meaning that many of the fields are managed lists of keywords, thesauri, ontologies and vocabularies.

To ensure consistency, different archival materials are catalogued using the same controlled vocabularies.

The system's linking capabilities enable maintaining different types of relationships between items of all kinds within the system. It allows sustaining an archival hierarchy, i.e. linking between the levels of collections, sub-collections, series, files and items. It also allows linking of a different kind: linking between records and digital representations of documents.

The software provides multilingual support which allows content management in parallel languages. Thus the cataloguers can produce multilingual descriptions of archival materials and name records. As a result, most of Yad Vashem's descriptions are accessible in several languages, mostly - but not exclusively - in English and Hebrew. This practice is quite distinct from the classical archival approach where the description of the material should be in the same language as the documentation itself.

Controlled vocabularies such as place names, keywords, ontologies of given and surnames are determined and managed by subject experts. In each vocabulary, a relationship between terms in a set is defined and adhered to. For each set of terms, a preferred term is selected and referenced by its variants and synonyms.

1 IDEA ALM system, https://idea-alm.com/about-us/

For example, ontologies of names, both first and surnames, are managed in such a way that allows maintaining an ever growing lexicon of variants. Name variants are clustered together in a set of synonyms. This process is led by onomastics specialists in the Archive's employ. The same applies also to place names and other keywords.

The Challenge

Over the past decades, the archival universe has been evolving and becoming increasingly more digital. Catalogues have become electronic, databases have been created. Finally, the documents themselves started to undergo digitization. Due to this advancement, the acquisition of digital materials by Yad Vashem has grown considerably – be it catalogues and databases or scanned documents. This significant influx in acquisition of digital resources brought about the recognition that despite all the considerable capabilities our knowledge management system allowed for, the Archive still lacked a way to process that new kind of data. This realization compelled the Archive to look for an approach that could enable it to deal with this type of material, with the main objective being to devise a way that would grant the end-users access to retrievable records presented in Yad Vashem's context, in a quick, timesaving manner and utilizing technologically savvy solutions.

For that purpose, a Data Integration Section was established. The section was tasked to convert the received data in order for it to be successfully ingested by Yad Vashem's cataloguing system. Since its creation, the Data Integration Section has handled many diverse projects, including name databases and lists, archival collections, testimonies collections and more.

The main difficulty, which became apparent virtually from the start, was a complete lack of uniformity between the different data sets - neither the descriptive materials nor the digital copies were arranged the same way. Each institution implemented different policies and procedures to digitize its documents. Diverse cataloguing software produced records of different types. The cataloging conventions themselves were inherently distinct in each project: the terminology, the keywords and their usage, and even the languages weren't the same. All those difficulties were further heightened by Yad Vashem's own cataloguing software with its protocols and controlled indexing procedures.

Data Integration Workflow

Over the years, in order to be able to import the received data harmoniously into the institution's cataloging system, the Section devised certain approaches, strategies and methods.

The data integration process was divided into four main stages: preliminary analysis, data model creation, data normalization, and final compilation.

The first stage, preliminary analysis, consists of two tasks: the gathering of background information, which consists of the historical and archival contexts, and, most importantly, the methodological framework utilized to produce the digital resources in question. Knowing the

historical setting in which certain documents were created and by who, and how a collection was shaped, arranged and re-arranged in an archive, all contribute to a better understanding of the resource and to its interpretation. The second task, understanding the methodology behind the digital resource, is of great importance to a successful project. Which conventions, standards, decisions and allowances guided the creation of the digital resource? What sort of metadata was chosen for cataloguing and how was it done? How was the digitization of documents carried out? Having answers to all of these questions is instrumental in order to correctly analyze, interpret and normalize the materials. Another aspect of the second task is the preliminary structural analysis of the received materials, including both the descriptions and the digitized documents. The way of arrangement of the metadata and of the digitized document images is investigated. Identifying the different metadata templates used in a given project allows bringing all data sets to a uniform, consistent structure for further analysis and subsequent modifications. Knowing the structure in which the image files were delivered is crucial for successfully establishing the correct relationship between the components of a digital resource. In order to achieve that it is important to understand how the image files are organized in a directory system, to find out whether there are metadata stored in attributes of those files or, whether some structural meaning can be gleaned from file names.

At the second stage an output data model is defined and according to it the project guidelines are finalized. The guidelines specify how a given collection should be arranged within Yad Vashem's data structure. They also state which controlled vocabularies should be used and list the fields that should be mapped and interpreted. A completion of this stage would be virtually impossible without the knowledge that was gathered at the first stage of the workflow.

The subsequent third stage, data normalization, is the most extensive and demanding. Here, again, the background information about the methodology behind the production of the metadata is vital. During this stage most of the data interpretation, cleaning and normalization is performed. At this stage we also look up and validate matches in the system's many controlled vocabularies. In some cases it is possible to expand Yad Vashem's vocabularies by adding new terms found in a new resource. Those new terms undergo an evaluation by a subject expert and are then added automatically to the system during data ingest. For example, when a new surname or a place name is detected within the metadata in question it will be sent for approval to the expert and after the expert's validation will be added to the appropriate vocabulary. In many cases during the course of one project the controlled vocabularies can be augmented by hundreds and sometimes thousands of new values.

The last stage is the compilation of the normalized data. The output tables, ready for ingestion, are produced. At this point it's possible to enhance the reworked metadata by adding multilingual (Hebrew and English) titles and enriching it with Yad Vashem's thesauri and keywords. The last task at this stage is the linking between restructured data and matching images which is made possible due to the file directory structure analysis that was performed during the first stage of the project.

Use Case: Project Latvia

The four stage workflow will be demonstrated by showcasing a project that the Data Integration Section carried out.

In 2015 the team took on a project to integrate archival materials and name lists received from Latvian collection holders.

Introduction

Yad Vashem acquired archival materials from Latvian State Historical Archives (Latvijas Valsts vēstures arhīvs, LVVA) and from the Latvian State Archives (Latvijas Valsts arhīvs, LVA). The materials included thirty-three microfilmed collections, among them collections from pre-war Jewish organizations, population censuses conducted in 1935 and 1941,[2] Riga House Committee books,[3] documentation regarding Jewish education originating from governmental and communal authorities and a collection of personal files of Jewish residents deported from Latvia by the Soviet authorities in June 1941.[4]

Yad Vashem also acquired a collection of domestic and foreign passports[5] that was delivered in the form of digitized document images,

In addition, Yad Vashem received a set of lists of names of Jewish residents from Latvia made by the Center for Judaic Studies at the University of Latvia.[6] The lists are the result of a research project that aimed to determine the fates of Latvian Jewry in the Holocaust. Some of the archival sources that were used to compile these lists were later acquired by Yad Vashem and are mentioned above.

Extent

The microfilmed collections were delivered to Yad Vashem on 900 reels.

The digitized passport collection amounted to 187,000 image files.

Descriptive metadata were obtained with the copied materials. The collections were accompanied by file level descriptions. Descriptions for both the microfilmed and the digitized collections were produced in 50 Excel files and 2 Word documents. In addition, there were three Excel files with lists of 90 thousand names of the Jewish residents of Latvia.

2 F.1308. Valsts Statistiskā pārvalde (Rīga), Latvian State Historical Archives (Latvijas Valsts vēstures arhīvs, LVVA).

3 F.2942. Rīgas pilsētas un apriņķa mājas grāmatas. LVVA.

4 F.1987. 1941.gada 14.jūnijā no Latvijas izsūtīto iedzīvotāju personas lietas, Latvian State Archives (Latvijas Valsts arhīvs, LVA).

5 F.2996. Rīgas Prefektūras pasu lietu kolekcija. LVVA.

6 This project was published on the Internet by the researchers as "The Latvian Names Project" in 2006. Since then it has been constantly updated and fine-tuned. http://names.lu.lv/

Each of these datasets is discussed separately below.

Objective

The thought behind the Latvian project was to produce as full a picture as possible of Latvian Jewish community before the war and its fate after the Holocaust. In order to achieve this, the project was split into four assignments:

1. To integrate the list of names of Jewish residents of Latvia into the Yad Vashem names database.
2. To integrate the archival descriptions of the copy collections into the Archive catalogue.
3. To extract names data from the metadata in archival descriptions and produce name records in the Yad Vashem names database.
4. To link between name occurrences within the scope of the project in order to be able to present a personal history of sorts based on archival sources as well as on the external research project ('List of Jewish residents').

Process

Preliminary analysis

Historical and archival context of the acquired collections

Looking into the historical and archival context of the collections in question allowed us to learn their essence, to understand who were their creators, what types of materials they contained, and whether they were homogenous or diverse in nature. This sort of information was vital later in determining the appropriate schema and templates for integration.

Microfilmed collections

Several collections were identified as being homogenous. For example, there was a set of personal files of Jewish teachers that was created by a special department at the Latvian Ministry of education that was solely dedicated to the subject of Jewish education[7]. The files, for the most part, consisted of similar document types. Another example of a uniform collection was the collection of 1935 Latvian population census[8]. The entire collection consisted of personal questionnaires filled out by census enumerators in a given locality. In contrast, most of the collections created by Jewish organizations were very diverse. They contained financial documents, lists of members, and correspondence with other branches and organizations around the globe, from Mandatory Palestine to Shanghai.

7 F.1632-1. Izglītības ministrija (Rīga), LVVA.

8 F.1308-12. Valsts Statistiskā pārvalde (Rīga), 1935. LVVA.

Collection of domestic and foreign passports

The original collection[9] in the Latvian Historical Archives contains about 500 thousand archival files, each in a separate envelope corresponding to a single person. The collection consists of applications to the Riga authorities to issue a new passport, usually to replace an expired one. In addition to a passport, an envelope may contain other personal documents (birth and marriage records etc.). Yad Vashem acquired digital copies of 43 thousand files from this collection. However, during the digitization only the passports were copied, without the accompanying documents. On top of that, when an envelope contained several passports, only one of them was digitized.

The information about the arrangement of the materials and the way they were selected for digitization proved to be very important in sorting and then linking the images to the descriptions. Since we knew that each envelope (file) could contain images pertaining only to one person, it helped validating results during special case checks – checks we perform before authorizing linking of image files to the descriptions.

Research the methodological framework used to produce data sets

The goal of this task was to gather as much information as possible on the standards adopted by the creators of the data sets, the policies and procedures implemented by them, and the assumptions and allowances that were made during the production and cataloguing of the data. This essential step allowed for more precise and accurate interpretation of the data for subsequent steps.

For example, while exploring the ways and the means used to produce the 'List of Jewish residents'[10] we learned that the team working on the project made a decision to ignore the original Latvian spellings of names and to transcribe all names to conform to a German spelling system. For example, the Latvian surname variant 'Broščermanis' was transformed into 'Broschtschermann'. This information proved to be useful first when we looked up matches for those surnames in Yad Vashem's ontology. It was especially important in cases when a proper match wasn't found and we needed to send those values for evaluation by an onomastics expert before adding them to the vocabulary. Second, being aware of this modification assisted us in plotting a strategy to link between different mentions about the same person within the project. So, when the same person was mentioned both in the 'List of Jewish residents' in a Germanized fashion and in the passports collection in proper Latvian form we could adjust for that and recognize the similarity.

Another example of a policy implemented by the authors of the List is a decision to spell place names according to the historical period of the event denoted. For all the events that occurred before 1919,[11] a non-Latvian variant was used, if existed (ex. Dünaburg or Dvinsk and not

9 F. 2996. Rīgas Prefektūras pasu lietu kolekcija. LVVA.

10 Most of the information was gathered at the project's website (http://names.lu.lv/en.html#meto). Additionally, some clarifications were obtained thanks to personal communication with the authors.

11 Renaming of Latvian cities and towns took place after the declaration of independence following the

Daugavpils). Here again, this knowledge helped us to validate the results after we matched all the place names to our vocabulary.

Understanding the terminology used to describe the fates of Jews was vital and enabled a proper interpretation of these keywords. For example, when a fate of a person was denoted as 'deported' it meant deportation to Siberia by the Soviet authorities in June 1941 and not deportation by the Nazis, which might have been the default thinking in the context of Yad Vashem.

Structural analysis of the data sets

At this step, the main emphasis was on metadata assessment. In order to proceed to the next stages of the project it was essential to identify the different templates of metadata.

Thus, by examining file formats, counting fields and comparing field names we were able to determine how many templates were used.

List of Jewish residents

The list was arranged in three excel files – two for residents of Riga and one for the residents of all other localities. Although there were apparent similarities, each of these lists used a separate template. The number of fields was different and the names of the fields varied. For example, fields named [Prewar Residence] and [Prewar Address] in one table were named [Res. Pre-WWII] and [Address Pre-WWII] in another. After additional analysis it became apparent that those three templates could be reduced to a single one.

Microfilmed collections

Within this data set of 50 files describing thirty-three archival collections, only one template was found. The only inconsistency that was apparent was due to different file formats of two collection descriptions, which were produced as formatted tables in Word documents.

Collection of domestic and foreign passports

This collection was delivered to Yad Vashem in two installments. Each installment was accompanied by an archival inventory list that constituted yet another template.

To sum up, by the end of this step it was revealed that all 55 different metadata files could be reduced to three templates.

Once these templates were recognized, three structurally cohesive data sets were prepared, one for thirty-three microfilmed collections, another for the passport collection and one more for the 'List of Jewish residents'.

Treaty of Brest-Litovsk between Russia and Germany in March 1918.

Output Data Model and Project Guidelines

After the first three preliminary steps were completed, the path was clear to select and fit existing Yad Vashem compatible schemas for the data sets. Two separate schemas were chosen for name indexing and three for cataloguing archival descriptions. Name records that emanated from archival descriptions were indexed in a different schema from the 'List of Jewish residents' because in addition to the biographical data, pointers to digital images could also be provided in the first case and not in the second. The archival collections were classified into three groups: personal files, official documentation (created by governmental entities) and nonspecific documentation (mostly created by Jewish organization). According to this classification three different schemas deemed appropriate.

The knowledge of historical and archival context (see above, Historical and archival context of the acquired collections) aided us in determining the proper schemas, especially for archival descriptions.

Following, final guidelines for the completion of the project were formulated.

Data Normalization

Mapping

This step was the first in a series of steps we took to normalize, clean and validate the data sets.

The fields were mapped to the selected schemas. These mappings determined which processes and steps needed to be taken next for each of the fields. Naturally, fields that were mapped to Yad Vashem controlled vocabularies were to be processed differently from date and text fields, for example.

During this step, we heavily relied on the information that was gathered in the beginning of the project (see, Preliminary analysis).

Data cleaning

In order for the data to conform to Yad Vashem standards, in most cases, a process of cleaning is needed. Most of these processes were written into automated routines and procedures.

In order to apply those procedures effectively, similar fields were grouped together. This enabled us to work with similar types of data simultaneously in order to standardize them.

For example, all the data mapped as date fields were grouped into a single list. Following that, a procedure was run to transform all the dates into patterns. The patterns were then analyzed using a preexisting list of interpreted patterns. Each new pattern was deciphered and added to the list for future reference. As a result, all the dates were formatted and propagated. (See Figure 1: the column [Dates] contains original data while the other columns are cleaned, formatted and interpreted).

ID	Dates	DATE_DEATH_	DATE_DEAT	STATUS_
9	1942 16 11	16/11/1942	16/11/1942	
28	1944 31 12	31/12/1944	31/12/1944	
31	1941	1941	01/01/1941	
32	1944 02 12, Still alive 1943 22 04	02/12/1944	02/12/1944	
33	1944	1944	01/01/1944	
34	1941 04 06	04/06/1941	04/06/1941	
36	1952 20 07	20/07/1952	20/07/1952	Survived
42	1941	1941	01/01/1941	
43	1941	1941	01/01/1941	

Figure 1: Formatted dates

In the case of fields mapped with table values, such as names, surnames or places, after grouping them together the data was normalized, split into separate fields if needed and adjusted to enable successful matching to the vocabularies. In many cases, data fields in the source material contained several values which needed to be separated in order to be matched to vocabularies. This was the case especially in the 'List of Jewish residents'. The field for surnames could contain a compound name. The field for place of residence could hold several names of localities (in cases when the researcher encountered several sources with different addresses). All those cases had to be recognized and evaluated. (See Figure 2: the column [Value] contains the input data, the column [StrMod] contains the separated substrings while the columns [Count] and [Index] denote their amount and order in the original).

Compilation

Finally, after all the cleaning and normalization was completed, all the appropriate matches were looked up and validated in the system's many controlled vocabularies, a new table, according to the selected schema, was compiled in order to integrate the data into the Yad Vashem catalogue.

Value	Count	Index	StrMod
DLIN-RIWOSCH	2	1	Dlin
DLIN-RIWOSCH	2	2	Riwosch
DRASNIN aka DRASNIK	2	1	Drasnin
EDELSTEIN-KERENI	2	1	Edelstein
EDELSTEIN-KERENI	2	2	Kereni
FISCHER-KAHN	2	1	Fischer
FISCHER-KAHN	2	2	Kahn
FRIEDMANN-RABINOWITSCH	2	1	Friedmann
FRIEDMANN-RABINOWITSCH	2	2	Rabinowitsch

Figure 2: Surnames after splitting into separate fields

As a part of this compilation step, some metadata enhancements were incorporated.

One of these enhancements was formulating descriptive titles at the archival file level. This was made possible by analyzing the original file titles and texts (mostly in Russian) in order to define patterns. As a result of this analysis about 150 different textual patterns were recognized. The patterns then were translated into English and Hebrew. Based on the translated patterns proper titles were calculated for each file. (See Figure 3: the first row represents an original Russian title, the second – a pattern, the last two rows demonstrate calculated titles in English).

Finally, all the restructured data (except 'List of Jewish residents')

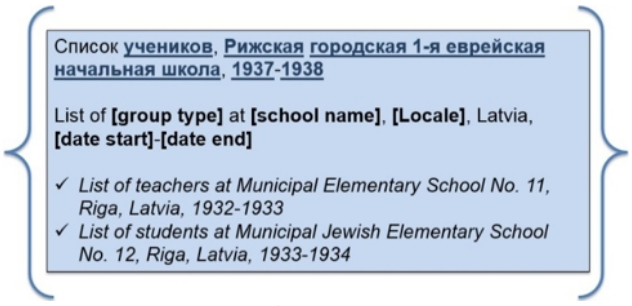

Figure 3: From patterns to titles

were linked to matching digital objects, i.e. digitized microfilm reels and scans.

All 900 microfilm reels underwent digitization after delivery to Yad Vashem. After this was completed it was possible to establish a relationship between ranges of image files in each reel and the appropriate descriptions. That allowed us to create pointers to pages in a microfilm reel.

As to the collection of passports, which was received as a digital copy, after evaluating the directory structure and file naming patterns, and eliminating duplicate images, an alternative file structure was proposed. The source files were copied and converted to conform to Yad Vashem's standards. Those derivative files were renamed and reorganized in a different file directory system. This reordering enabled proper linking between the two components of the archival file: the descriptions and the images.

Results

At the completion of the project, the following records were integrated into the Yad Vashem database:

1. 90,000 name records from List of Jewish residents in Latvia
2. 10,291 files of microfilmed copy collections
3. 43,000 files of passports collection
4. 332,000 name records extracted from metadata in archival descriptions
5. 86,400 groups of name occurrences were clustered together.

Blanket approach?

Adhering to this four stage process enabled proper integration into the main catalogue of new, retrievable records, which were consistent with the rest of Yad Vashem's collections. This also provided a variety of entry points to the catalogue that made the records in question even more accessible for the end-users in-house as well as worldwide.

Since it is applicable to any data integration project the Data Integration Section takes on, the epithet "blanket approach" deemed appropriate.

Applying the four stage workflow has proven itself efficient in other projects so far as well.

The utility of this approach may, at least in part, be due to the conservatism of the domain. After all, we have been consistently dealing with materials that lack uniformity, as the aspect of potential data sharing doesn't seem to be taken into account when various organizations take on cataloguing and indexing projects.

Conclusions

Overall, the Data Integration Section's work can be viewed as a success; since its launch it enabled the integration of nearly two million name record and almost 400 thousand archival descriptions.

The extensive analysis that is done during the work on a specific project provides opportunities to assist other departments on the collections in question. Among those is the Archival Acquisitions Department, which can obtain the Section's feedback that may be beneficial for future procurements.

However, it seems prudent to call for better and more widespread adoption and implementation of international standards. The lack thereof can be the chief reason for two main shortcomings of the current process – the risk of overgeneralization and of misinterpretation.

The adoption of these standards would allow investing less effort in formal data interpretation and transformation, and more in contextual interpretation, which will increase the likelihood to provide better and richer descriptions.

Last URLs consultation: 2019, February, 3.