Semantic Archive Integration for Holocaust Research. The EHRI Research Infrastructure

[1]Vladmir Alexiev, [2]Ivelina Nikolova and [3]Neli Hateva

Ontotext Corp., Sofia, Bulgaria
[1]vladimir.alexiev@ontotext.com
[2]ivelina.nikolova@ontotext.com
[3]neli.hateva@ontotext.com

Abstract. The European Holocaust Research Infrastructure (EHRI) is a large-scale EU project that involves 23 institutions and archives working on Holocaust studies, from Europe, Israel and the US. In its first phase (2011-2015) it aggregated archival descriptions and materials on a large scale and built a Virtual Research Environment (portal) for Holocaust researchers based on a graph database. In its second phase (2015-2019), EHRI-2 seeks to enhance the gathered materials using semantic approaches: enrichment, co-referencing, interlinking. Semantic integration involves four of the 14 EHRI-2 work packages and helps integrate databases, free text, and metadata to interconnect historical entities (people, organizations, places, historic events) and create networks. We will present some of the EHRI-2 technical work, including critical issues we have encountered.

EHRI – European Holocaust Research Infrastructure – è un progetto europeo cui partecipano 23 fra istituti di ricerca, enti di conservazione e aziende informatiche in Europa, Israele e Stati Uniti. Nella prima fase del progetto (EHRI-1, 2011-2015) è stata avviata una raccolta su larga scala di descrizioni archivistiche e materiali sulla Shoah che sono state integrate nel Virtual Research Environment (EHRI Web Portal) basato su Graph DB. Nella seconda fase del progetto (EHRI-2, 2015-2019) si sta cercando di valorizzare i materiali raccolti utilizzando approci di tipo semantico (enrichment, interlinking, co-referencing). In questa attività sono coinvolti quattro work-packages (sui 14 dell'intero progetto), tutti impegnati per l'integrazione di database, testi e metadati al fine di connettere fra di loro entità varie (persone, enti, luoghi, eventi storici) e creare così dei network di conoscenza. In questo paper vengono presentate le attività svolte dai vari workpackages di EHRI 2 per l'integrazione dei dati, dando spazio e rilievo anche alle criticità incontrate nel corso del lavoro.

Introduction

The European Holocaust Research Infrastructure (EHRI) is a large-scale EU project that involves 23 institutions working on Holocaust studies, from Europe, Israel and the US. In its first phase (2011-2015) EHRI aggregated archival descriptions and materials on a large scale and built a Virtual Research Environment (https://portal.ehri-project.eu/) for Holocaust researchers based on a graph database (neo4j) 3.. EHRI results were presented in several Holocaust-related magazines 2.; 10. and conferences 5.; 14..

In its second phase (2015-2019), EHRI-2 seeks to enhance the gathered materials using semantic approaches: enrichment, co-referencing, interlinking, geo-mapping, named entity recognition, topic extraction and mapping, etc. Four of the 14 EHRI-2 work packages (WP10, WP11, WP13, WP14) use Semantic Integration approaches, which helps integrate databases, free text, and metadata to interconnect historical entities (people, organizations, places, historic events) and create networks. In detail:

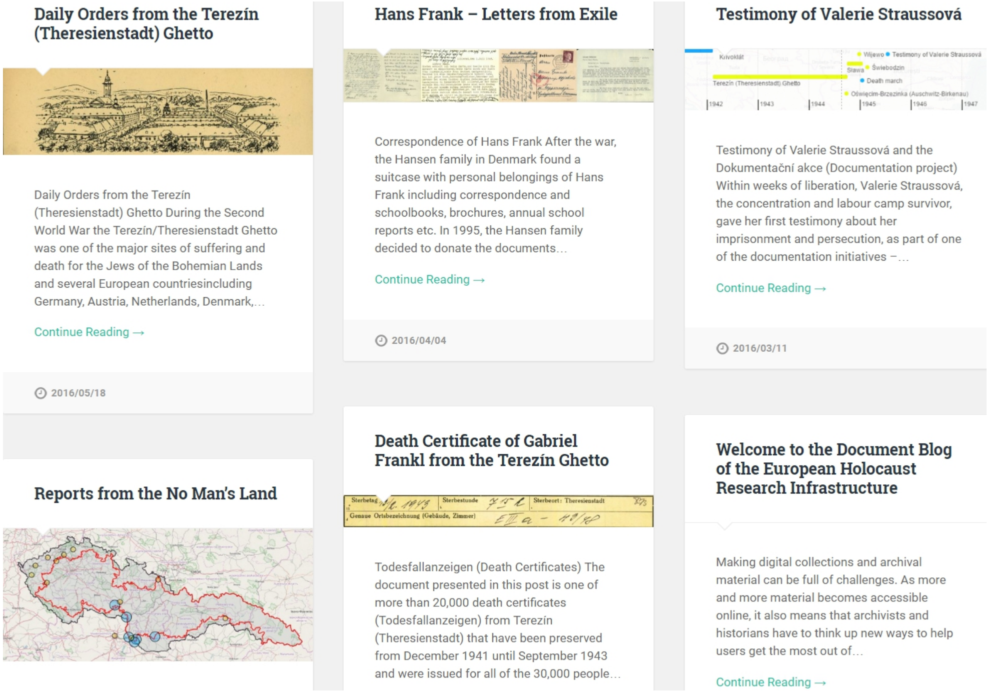
- **WP10 (EAD)** converts archival descriptions from various formats to standard EAD XML; transports EADs using OAI PMH or ResourceSync; ingests EADs to the EHRI database; enables use cases such as synchronization; co-referencing of textual Access Points to proper thesaurus references.
- **WP11 (Authorities and Standards)** consolidates and enlarges the EHRI authorities to render the indexing and retrieval of information more effective. It addresses Access Points in ingested EADs (normalization of Unicode, spelling, punctuation; deduplication; clustering; co-referencing to authority control), Subjects (deployment of a Thesaurus Management System in support of the EHRI Thesaurus Editorial Board), Places (co-referencing to Geonames); Camps and Ghettos (integrating data with Wikidata); Persons, Corporate Bodies (using USHMM HSV and VIAF); semantic (conceptual) search including hierarchical query expansion; interconnectivity of archival descriptions; permanent URLs; metadata quality; EAD RelaxNG and Schematron schemas and validation, etc.
- **WP13 (Data Infrastructures)** builds up domain knowledge bases from institutional databases by using deduplication, semantic data integration, semantic text analysis. It provides the foundation for research use cases on Jewish Social Networks and their impact on the chance of survival.
- **WP14 (Digital Historiography Research)** works on semantic text analysis (semantic enrichment), text and entity similarity, geo-mapping. It develops Digital Historiography researcher tools, including Prosopographical approaches.

First we present two examples of using semantic linking (the EHRI Blog and CDEC collection), which provide motivation for part of the EHRI technical work. Then we present the work packages. We conclude with a summary, lessons learned and outstanding challenges.

The EHRI Document Blog

The EHRI Document Blog (https://blog.ehri-project.eu/) was started as a space to share ideas about Holocaust-related archival documents, and their presentation and interpretation using digital tools. EHRI researchers document their activities and experiment with different ways to explain and show digital archival content. The blog is a showcase of using novel approaches for digital archival research.

The blog also provides inspiration to the technical work packages (described below) as to what functionalities and analyses can be useful to researchers, in order to automate part of their work, and allow them to analyze such large amounts of data that was not possible previously.



Daily Orders from the Terezín (Theresienstadt) Ghetto

Daily Orders from the Terezín (Theresienstadt) Ghetto During the Second World War the Terezín/Theresienstadt Ghetto was one of the major sites of suffering and death for the Jews of the Bohemian Lands and several European countriesincluding Germany, Austria, Netherlands, Denmark,…

Continue Reading →

2016/05/18

Hans Frank – Letters from Exile

Correspondence of Hans Frank After the war, the Hansen family in Denmark found a suitcase with personal belongings of Hans Frank including correspondence and schoolbooks, brochures, annual school reports etc. In 1995, the Hansen family decided to donate the documents…

Continue Reading →

2016/04/04

Testimony of Valerie Straussová

Křivoklát
Wijewo
Testimony of Valerie Straussová
Sława
Świebodzin
Death march
Terezín (Theresienstadt) Ghetto
Oświęcim-Brzezinka (Auschwitz-Birkenau)
1942
1943
1944
1945
1946
1947

Testimony of Valerie Straussová and the Dokumentační akce (Documentation project) Within weeks of liberation, Valerie Straussová, the concentration and labour camp survivor, gave her first testimony about her imprisonment and persecution, as part of one of the documentation initiatives –…

Continue Reading →

2016/03/11

Reports from the No Man's Land

Death Certificate of Gabriel Frankl from the Terezín Ghetto

Sterbetag
Sterbestunde
Sterbeort: Theresienstadt
Genaue Ortsbezeichnung (Gebäude, Zimmer)

Todesfallanzeigen (Death Certificates) The document presented in this post is one of more than 20,000 death certificates (Todesfallanzeigen) from Terezín (Theresienstadt) that have been preserved from December 1941 until September 1943 and were issued for all of the 30,000 people…

Welcome to the Document Blog of the European Holocaust Research Infrastructure

Making digital collections and archival material can be full of challenges. As more and more material becomes accessible online, it also means that archivists and historians have to think up new ways to help users get the most out of…

Continue Reading →

Picture 1: EHRI Document Blog. Courtesy EHRI 2016

The Value of Linking

An early inspiration that helped convince the EHRI Project Management Board of the value of interlinking, was the work performed by the Contemporary Jewish Documentation

Center Foundation (CDEC), Milan, in linking their internal databases (archival materials and persons) to Linked Open Data (LOD).

For example, one of the 9k person records that CDEC has is about Primo Levi, an Italian Jewish chemist, memoirist, short story writer, novelist, essayist, and Auschwitz survivor. This person is present in 50 wikipedias (articles about him, e.g. see Primo Levi on en.wikipedia[1]), 11 wikiquote sites (his sayings), and 2 wikisource sites (his books).

There is quite a lot of structured Linked Open Data (LOD) about him: Primo Levi on wikidata[2] (or see Primo Levi on Reasonator,[3] which is a nice reading interface):

- Names in a number of languages (including Russian, Chinese, Korean, Arabic): Primo Lévi | Primo Michele Levi | Primo levi | Primo M. Levi | Levi, etc.
- His prisoner number (174517)
- Identifiers in a large number (30) of national library systems and other sources, i.e. cross-referencing to the following Authority Files BAV, BNE, BnF, British Museum person-institution, CANTIC-ID, DBNL, FAST-ID, Filmportal, Find a Grave grave, Freebase, GND, IMDb, ISNI, KLfG Critical Dictionary of foreign contemporary literature, LCAuth, Munzinger IBA, NDL, NILF author, NKCR AUT, NUKAT authorities, National Library of Israel, National Thesaurus for Author Names, Open Library, PTBNP, Perlentaucher, RKDartists, SBN, SELIBR, SUDOC authorities, VIAF. Curiously, he has 5 identifiers in the National Library of Israel and 3 ISNI identifiers. Additional info can be obtained from these sources, especially publications.
- Auto-generated "biography" (rather dry): Primo Levi was an Italian chemist, poet, politician, writer, and novelist. He played a role in imprisonment at Auschwitz concentration camp. He was born on July 31, 1919 in Turin. He died on April 11, 1987 in Turin.
- Nationality & professions, each supported by a primary source URL.
- 19 books that he authored
- Prizes he has won
- Who he was influenced by, and who he influenced
- Familial relations, if available

One of the external sources, the International Movie DataBase (Primo Levi on IMDB),[4] shows that he has credits in 11 films, news about him (5 in 2015 alone), and a longer biography.[5]

[1] https://en.wikipedia.org/wiki/Primo_Levi

[2] https://www.wikidata.org/wiki/Q153670

[3] https://tools.wmflabs.org/reasonator/?q=Q153670&lang=en

[4] http://www.imdb.com/name/nm0505485/

[5] http://www.imdb.com/name/nm0505485/bio?ref_=nm_ov_bio_sm

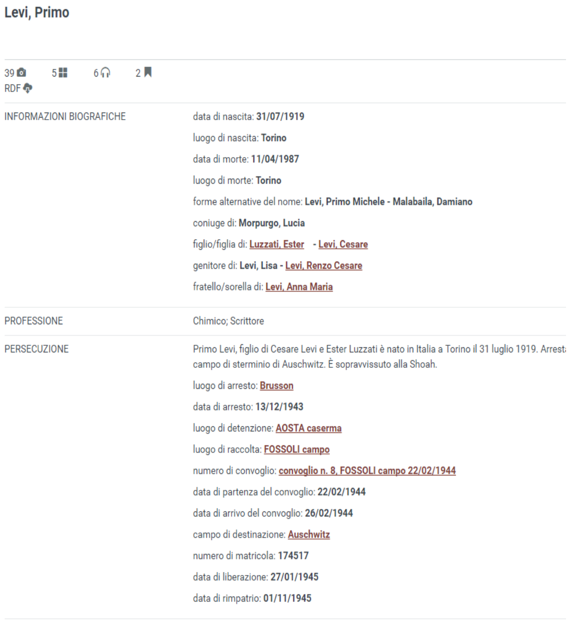
IMDB is also available as LOD, see LinkedMDB.[6]

CDEC Linked Data

CDEC has connected person names to archival descriptions, the CDEC library catalogue, and external LOD (persons to VIAF and DBpedia, places to Geonames). So when a user searches for a person, they find biographical data, family relations, information about the available resources on that person at the CDEC archives and library. For example, http://digital-library.cdec.it/cdec-web/persone/detail/person-5002/levi-primo.html shows:

- Links to 39 photos, 5 archival materials, 6 audio files, 2 bibliographic items
- Link to get LOD data in RDF: http://dati.cdec.it/lod/shoah/person/5002 using the Italian Shoah ontology 12.. Thus CDEC names and biographical data are themselves published as LOD
- Structured properties about the person and interlinking through familial relations

Picture 2: Primo Levi on CDEC Catalog. Courtesy CDEC, 2016



Levi, Primo

39 📷 5 ▦ 6 🎧 2 🔖
RDF

INFORMAZIONI BIOGRAFICHE	data di nascita: **31/07/1919** luogo di nascita: **Torino** data di morte: **11/04/1987** luogo di morte: **Torino** forme alternative del nome: **Levi, Primo Michele - Malabaila, Damiano** coniuge di: **Morpurgo, Lucia** figlio/figlia di: **Luzzati, Ester** - **Levi, Cesare** genitore di: **Levi, Lisa** - **Levi, Renzo Cesare** fratello/sorella di: **Levi, Anna Maria**
PROFESSIONE	Chimico; Scrittore
PERSECUZIONE	Primo Levi, figlio di Cesare Levi e Ester Luzzati è nato in Italia a Torino il 31 luglio 1919. Arrest campo di sterminio di Auschwitz. È sopravvissuto alla Shoah. luogo di arresto: **Brusson** data di arresto: **13/12/1943** luogo di detenzione: **AOSTA caserma** luogo di raccolta: **FOSSOLI campo** numero di convoglio: **convoglio n. 8, FOSSOLI campo 22/02/1944** data di partenza del convoglio: **22/02/1944** data di arrivo del convoglio: **26/02/1944** campo di destinazione: **Auschwitz** numero di matricola: **174517** data di liberazione: **27/01/1945** data di rimpatrio: **01/11/1945**

[6] http://www.linkedmdb.org/

The RDF data is also published on a SPARQL endpoint (http://lod.xdams.org/sparql) and can be queried, e.g. "persons with profession Chemist" or "persons with destination camp Auschwitz"

LOD linking allows CDEC to present a lot more data about the persons.

This work was performed as part of the Open Memory Project 4., which won the LODLAM Challenge 2015[7] for its significance in the virtual reunification of families.

Digital Historiography Research

WP14 (Digital Historiography Research) experiments with new technologies to develop Digital Historiography researcher tools, including Prosopographical approaches. These technologies include leveraging large semantic data sources (including LOD), semantic text analysis (semantic enrichment), text similarity (e.g. clustering based on Neural Networks, LDA), geo-mapping, etc.

EHRI Research Use Cases

One of the first tasks of WP14 was to define Research Use Cases that guide the work not only in WP14, but also technical development and capabilities in the other WPs. For example, which datasets to use as Geographic authorities in WP11 is dictated to some extent by needs identified in WP14. The cases were formulated by historians and EHRI archivists, together with digital humanities researchers and IT specialists. They are as follows:

1) **Names and Networks**: Holocaust victim communities. Most Jews needed the support of other people to survive. Persons found aiders inside their personal and group networks. Investigate the networks in which European Jews operated during their persecution in the Second World War, and improve understanding of the various chances of survival that persecuted Jews all over Europe had.

2) In Search of a Better Life and a Safe Haven: **Tracing the Paths of Jewish Refugees** (1933-1945). Map and better understand the different migration trajectories and determine the factors that played a role in the migration movements of migrants or forcefully deported Jews.

3) People on the Move: Revisiting **Events and Narratives of the European Refugee Crisis** (1930s-1950s). Investigate migration movements of European refugees. The International Tracing Service (ITS), and EHRI partner, has been a key actor in managing this twentieth century migration crisis and has hence built the most important archive documenting the lives of refugees and displaced persons.

4) Between Decision Making and Improvisation: Tracing and Explaining **Patterns of Prisoners' Transfers** through the Concentration Camp System. Investigate how the

[7] http://summit2015.lodlam.net/2015/04/21/challenge-entry-open-memory-project/

SS in the years 1942-1945, via Hollerith Departments and punch-card technology, managed the transport of inmates through the concentration camp system that covered the whole of occupied Europe

5) **Archives and Machine Learning**. Investigate how digital methods might support archivists in the creation of interoperable and consistent descriptions of sources (metadata) and in the linking of sources. Estimate metadata quality using data mining and machine learning approaches.

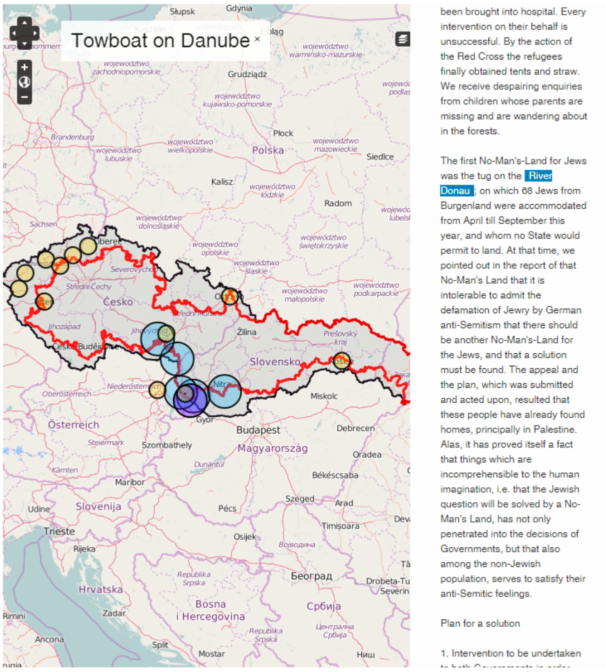
6) **Networked Reading**. What form historical sources need to take in order to be processed using digital methods? What kind of infrastructure do we need to extract meaning from them? How to apply digital methods in such a way that the results are verifiable as well as reproducible?

There is a lot of connections and cross-pollination between WPs. The same datasets, techniques and developed services can often be used in several WPs.

Geographic Mapping

Consider the following example from the EHRI Document Blog 9.. It presents an oral history about Jews stranded on the border of 3 countries around the Danube. An interactive map is displayed that shows the historical and current borders, the places being discussed, and links them between text and map (i.e. if you click on the map, you see the place name occurrences in the text, and vice versa).

Researchers find such visualizations and tools very useful, but creating them by hand is very effort intensive.



been brought into hospital. Every intervention on their behalf is unsuccessful. By the action of the Red Cross the refugees finally obtained tents and straw. We receive despairing enquiries from children whose parents are missing and are wandering about in the forests.

The first No-Man's-Land for Jews was the tug on the River Donau ; on which 68 Jews from Burgenland were accommodated from April till September this year, and whom no State would permit to land. At that time, we pointed out in the report of that No-Man's Land that it is intolerable to admit the defamation of Jewry by German anti-Semitism that there should be another No-Man's-Land for the Jews, and that a solution must be found. The appeal and the plan, which was submitted and acted upon, resulted that these people have already found homes, principally in Palestine. Alas, it has proved itself a fact that things which are incomprehensible to the human imagination, i.e. that the Jewish question will be solved by a No-Man's Land, has not only penetrated into the decisions of Governments, but that also among the non-Jewish population, serves to satisfy their anti-Semitic feelings.

Plan for a solution

1. Intervention to be undertaken

Picture 3: Reports from the No Mans Land. Courtesy Michal Frankl, Jewish Museum in Prague and EHRI, 2016

Geo-Referencing Service

EHRI has established a geo-referencing service that uses Geonames and DBpedia/Wikidata to find place references in text, access points (see WP11 below) and local databases (e.g. USHMM Places).

We use sophisticated place recognition pipelines developed by Ontotext based on commercial applications and extended to suit the specific EHRI domain needs. They make use of a hybrid approach combining machine learning algorithms and refinement rules trained on Gold Standards developed in accordance with detailed guidelines. For example, neither of these are places because they refer to organizations, institutions, papers, ships, etc.

- University of [Chicago], [Tel Aviv] university, Veterinary Institute of [Alma-Ata]
- The interview was given to the [United States] Holocaust Memorial Museum on Oct. 30, 1992
- in the Theater de [Champs Elysee]
- [Washington] Post, [Washington] Monthly
- USS [America]

Sophisticated disambiguation mechanisms which help to select the appropriate location from our knowledge base, in case more than one candidate exist, are developed. They are based on the following place characteristics:

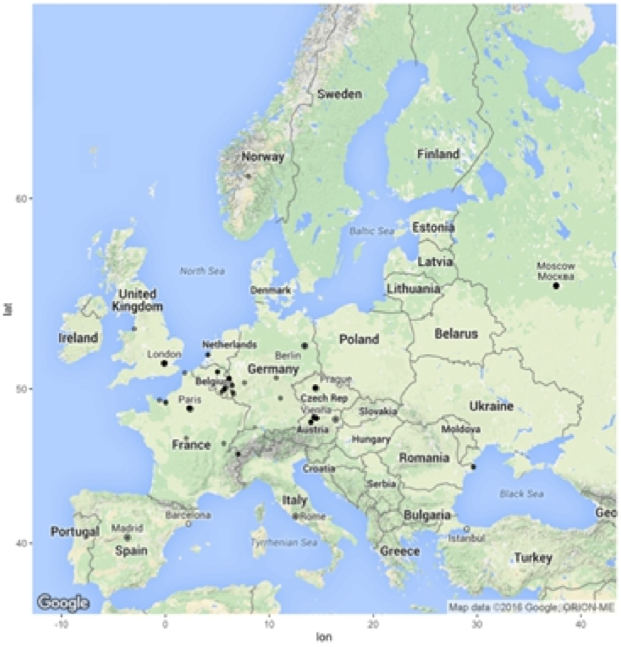
- Place names. We use a variety of synonym labels (Wikipedia redirects) and languages; e.g. *Oświęcim,* Освенцим, *Auschwitz, Auschwitz-Birkenau, Birkenau, Konzentrationslager Auschwitz, KZ Auschwitz.*
- Geonames feature types. E.g. a number of irrelevant place types (e.g. hotels) are removed.
- Place hierarchy. When a place is mentioned near its parent place, this allows us to disambiguate the child. This is especially important for Access Points, which often mention the place hierarchy, *e.g.*: Moscow (Russia): as opposed to the 23 places of the same name in the US, and a few more in other countries; Russia– Moscow; Alexandrovka, Lviv, Ukraine: "Alexandrovka" is a very popular village name. So although this doesn't lead to a single disambiguated place, it helps to reduce the set of possible instances from about 70 to about 20.
- Co-occurrence statistics based on GS over news corpora.
- Population: for populated places, we give priority to bigger places.
- Places nearer to Berlin are given priority.

An additional feature of the service is that if it finds several hierarchical places in one access point, it emits only the most specific places. Eg.:

- Moscow--Russia, Wilna--Poland: will be disambiguated to the following 3 places
- Moscow<Russia, Vilnius<Lithuania, Poland: see next section on the historic belonging of Vilnius

Geo-referenced place names are useful for various purposes:

- Geo-mapping of textual materials, as shown above.
- Other geographic visualizations, e.g. of detention/imprisonment vs liberation
- The place hierarchy can be used to extract records related to a particular territory, e.g. "Archival descriptions mentioning Ukraine" should find all records mentioning a place in Ukraine
- Coordinates can be used to map places, and compute distance between places
- Places are an important characteristic to consider when deduplicating person records. Certain probabilistic inferences can be made based on place hierarchy and proximity



Picture 4: Map of places extracted from a USHMM Oral History interview. Courtesy Tobias Blanke, Kings College London, 2016

.

Geographic Challenges

A significant challenge we faced is the (in)adequacy of LOD sources for historic geography. Nazis established their own "Nazi geography" when they occupied a country:

- Nazis renamed place names, e.g. Oświęcim->Auschwitz, Brzezinka->Birkenau.

- Nazis established administrative districts, e.g. Reichskommissariat Ostland[8] included Estonia, Latvia, Lithuania, the northeastern part of Poland and the west part of the Belarusian SSR

In addition, other historic processes changed borders, place names and administrative subordination. For example, Wilno (Vilna) was part of Poland until 1939, when USSR gave it to Lithuania, to become its capital Vilnius.

While Geonames often includes historic place names and even historic countries (e.g. Czechoslovakia), the Geonames place hierarchy reflects modern geography. Currently our geo-service uses this (e.g. Wilna--Poland is disambiguated as the two places Vilnius<Lithuania and Poland), but we have considered making local Geonames additions (e.g. make Czechoslovakia the parent of Czech Republic and Slovakia).

EHRI has a list of 261 Nazi administrative districts,[9] which is also available as SKOS at http://data.ehri-project.eu. However, these need to be linked to LOD (e.g. DBpedia) and the territorial hierarchy needs to be established. E.g.: https://portal.ehri-project.eu/keywords/admindistricts-212 (Reichsprotektorat in Böhmen und Mähren) should be linked to https://en.wikipedia.org/wiki/Protectorate_of_Bohemia_and_Moravia.

We are considering some other sources of historical geography, e.g. the Spatial History Project 7..

Jewish Social Networks

The "Names and Networks" research use case wants to investigate Jewish Social Networks and their impact on the chance of survival during the Holocaust. To this end we first need to build up person networks, which consists of (at least) two steps: (1) Deduplication/co-referencing of person records; (2) Rebuilding family and acquaintance relations from partial data. Both of these tasks involve probabilistic judgements, because of partial info (see the numbers under "*additional info*" below), and uncertain matching rules (e.g. approximate name matching, matching a name against an initial, approximate date matching, etc).

The following archival sources are considered for potential inclusion in the "Names and Networks" case:

- United States Holocaust Memorial Museum (USHMM) Persons (HSV) database (millions of names)
- Dutch Jewish Digital Monument website database (hundreds of thousands of names)
- CDEC person database (tens of thousands of names)
- Dutch Jewish Council proceedings (textual, not structured records)

[8] https://en.wikipedia.org/wiki/Reichskommissariat_Ostland

[9] https://portal.ehri-project.eu/vocabularies/admindistricts

- Virtual International Authority File (VIAF) (tens of millions of names, but only "famous" people, so potential little coverage of the Holocaust domain)
- Wikidata (couple million names, potential little coverage of the Holocaust domain)

For the time being we are actively working on USHMM Persons. USHMM Persons includes names about 3.2M **people** in a public part and another 3M in a private/secured part. For now we are working with the public part, but after elaborating specific security procedures, will also tackle the private part.

The public part also includes the following *additional info* (here we list only the most significant numbers):

- 1M additional **names** (392k patronymic, 233k mother's, 143k maiden, 105k father's, 68k spouse),
- 102k **family relations** to head of household,
- **Identification numbers** (folder/page/record/line, 147k prisoner identification, 142k age, 74k convoy, 49k depot, 53k transport identification, 42k number of people in transport)
- **Historic dates** (2.2M birth, 254k death, 78k arrival, 74k convoy, 55k departure, 50k transport, 19k liberation, 14k arrest, 10k marriage, 5.3k birth of spouse),
- **Historic places** (871k birth, 436k wartime, 359k residence, 215k death, 85k registered, 74k convoy destination).
- Categorical or descriptive **values** (501k occupation, 371k nationality, 16k religion, 116k marital status), 103k Holocaust fate, 53k ethnicity).

This provides rich historical info that can enable person identification (deduplication) and reconstructing person networks.

Of course, the data is not without quality problems. Name spelling is not always consistent; many pieces of info were not readable in the original materials (e.g. dates may have only month without year, or even only day without month); many records from different sources represent the same person, therefore deduplication (record linking) is required; place records also include duplicates, information is sometimes put in the wrong field, etc.

Historic Reasoning for Person Identification

In addition to well-known record linking (object identification) methods, we propose to use the family relations, temporal and geographic context of each record to facilitate probabilistic matching and linking.

But how exactly can we use the wealth of info in the various fields of the USHMM Person database? Which fields and what rules should we use, and what weights or probabilities should we assign to them? That is a challenging research question that we do not hope to resolve comprehensively, and are currently working based on examples.

We have selected 70 person name occurrences from USHMM Oral History interviews. We added the interviewee name for each one. We built a **Gold Standard** (GS) that tries to match these names to USHMM Persons using name search and historic reasoning.

The USHMM person records come from 5.1k **lists (sources)**, some of which correspond to **coherent historic events**. For example, source 20990[10] and catalog record irn518183[11] "Liste des Enfants avec Adresses":

- Describe a dated historic event that provides some probable info about each person in the list: that he/she was a **child**, was in **North Africa** at the time of making the list, was Considered for Emigration to the **United States**, and the record comes from the **Refugee Service, North Africa**.
- Provide evidence that it's likely the persons listed in that source knew each other.

Consider Oral History interview RG-50.469.0006 of Erwin Weissmann.[12] The selected GS sentence "Her maiden name was Stern, Malvine Stern" describes his mother. We learn from other places in the interview that his mother came from Tokai, Hungary; was married in 1905/6, moved to Baden bei Wien, Austria. She had three children: first one was born in 1907; second one (the interviewee) in 1909; and third one in 1911. The mother's maiden name was Stern and her married name was Weissmann. The interviewee (son) was deported in 1938 to Belgium.

From USHMM search we find that Malvin, Malvina, Malvine are all female names and thus represent variants of the same name. Malvin* Stern (maiden name) finds 17 records, and Malvin* Weissmann (married name) finds 3 records. (We use Exact and Fuzzy search but not Soundex search, which returns too many false positives). We have evaluated these 20 candidate records and concluded that neither is a likely match for the woman described in the interview.

E.g. consider Malvine Stern, PersonId=4225334:[13] she was born in 1870, so she would have had to be 35/6 years old when she got married, and gave birth to three children at the age of 37, 39 and 41, which is highly unlikely.

This record originates from the Assets Transfer Office of the Nazi-era Ministry of Commerce and Transportation. A decree concerning the Reporting of Jewish assets of April 26, 1938 required all Jewish citizens to report their total domestic and foreign assets, where such assets exceeded 50,000 Reichsmarks. So we know this candidate was wealthy and was harassed at some time after 1938, but this does not represent a coherent.

[10] http://www.ushmm.org/online/hsv/source_view.php?SourceId=20990

[11] http://collections.ushmm.org/search/catalog/irn518183

[12] http://collections.ushmm.org/oh_findingaids/RG-50.469.0006_trs_en.pdf

[13] https://www.ushmm.org/online/hsv/person_view.php?PersonId=4225334

From the above we can formulate a variety of rules regarding name matching (Malvin, Malvine and Malvina are variants of the same name), names and genders (see next), birth date of mother vs child (child bearing chance decreases sharply after 35), etc.

An example of another rule: we can assume that an arrest date would not be earlier than a person's 6th birthday, giving us a latest possible birth date for the person (if that is missing). Conversely, if we have two records that are considered possible matches, but the birth date in one disagrees (succeeds) with the arrest date in another, this is evidence against the match.

Place Reasoning

After co-referencing USHMM Places to Geonames (and deduplicating them as a result), we can use the Geonames place hierarchy, as well as place proximity (based on coordinates) to make some likely inferences.

E.g. consider this sentence from interview RG-50.661.0001:[14] "My Grandfather owned a bank in partnership with a cousin, Leibela Mandell". By using a search for Lejbela (First name, Soundex) and Mandel (Last name, exact) we find Lejb Ela MANDEL.[15] This record comes from a list Initial Registration of Lublin's Jews - October 1939 and January 1940,[16] described as "listing of the male heads of households appears to have survived in the Lublin Judenrat files". This list represents a coherent historic event/context: place=Lublin, dates=1939-10 to 1940-01, gender=male. To the untrained ear the name Lejbela seems female, but we learn from the interview that Leib is a male name: "second oldest son was Leib".

The interview also says "The family remained in Poland until the second world war", and mentions Premishlan (Przemyslany in Polish). In modern geography, Przemyslany does not have as ancestor place Poland, since the place is now Peremyshlyany, in Lviv Oblast, in Ukraine. However, we learn from Google Maps[17] that the distance from Lublin to Peremyshlyany is 263km. So it is quite possible that Lejb Ela MANDEL[18] is the same person as the cousin mentioned in the interview.

In addition, we will be able to find all person records related to any place in a certain country. This is useful to make sub-selections for particular research purposes, e.g. NIOD wants to investigate networks of Dutch Jews. USHMM has a list of 17k Dutch Jews, but we believe that we can extract a lot more Netherlands-related person records, perhaps up to 100k.

We also apply the geo-referencing service to help for deduplication of person records done under the "Names and Networks" research use case (described in 3.1). We use geo-referencing

[14] http://collections.ushmm.org/oh_findingaids/RG-50.661.0001_trs_en.pdf

[15] https://www.ushmm.org/online/hsv/person_view.php?PersonId=3120613

[16] https://www.ushmm.org/online/hsv/source_view.php?SourceId=20874

[17] https://www.google.bg/maps/dir/Lublin,+Poland/Peremyshlyany,+Lviv+Oblast,+Ukraine/@50.4566024,22.4417333,8z/data=!3m1!4b1!4m13!4m12!1m5!1m1!1s0x472257141e154061:0x5528ee7af6e8e95f!2m2!1d22.5684463!2d51.2464536!1m5!1m1!1s0x473a98e613b80eb1:0xe95d759ed61f558e!2m2!1d24.5593173!2d49.6692136?hl=en

[18] https://www.ushmm.org/online/hsv/person_view.php?PersonId=3120613

for location properties such as *place_of_birth* and *place_of_death* which allows us to have a single representation for places with alternative spellings and multiple names. By doing that we improve the usefulness of the feature representing whether two persons have the same place of birth or death and increases the chance to recognise whether 2 person records refer to the same person or not.

Reconstructing Families

Consider the following made-up example of data about personId 123456: firstName "John", lastName "Smith", gender Male, nameSpouseMaiden "Matienzo", nameSpouse "Maria Smith", dateMarriage 1921-01-05, nameChild "Mike Smith", nameSibling "Jack Jones".

We can create Person records for the additional names mentioned, make the following likely inferences, and then try to match them to other Person records in the database:

- The spouse (Maria) has gender Female and her maiden name is Matienzo
- They were married on the same date (in the same event)
- The child (Mike) has the person and his spouse respectively as Father and Mother
- The child's Birth date was likely after the Marriage date
- The sibling (Jack) is the child's uncle or aunt

Thus we can convert fielded **data** about persons into a **network of persons and events**.

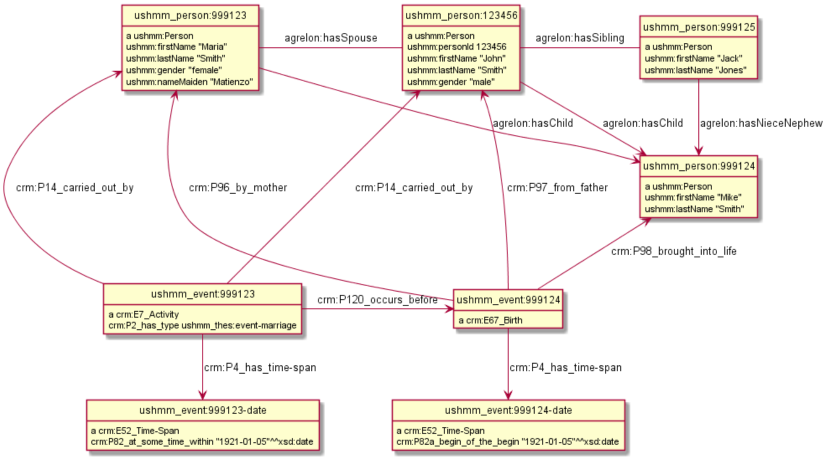


Illustration 5: Reconstructing person networks. Courtesy Vladimir Alexiev, Ontotext, 2016

We propose to use the **Conceptual Reference Model** (CIDOC CRM, standard ISO 21127:2014) ontology 3. to represent people and events, and the **Agent Relation Ontology** (AgRelOn) 13. to represent relations between people:

- The marriage is represented as crm:E7_Activity with crm:P2_has_type <marriage>, and the child's birth as crm:E67_Birth. The two events are linked with crm:P120_occurs_before (one of Allen's interval algebra relations). Participants are connected to the events using appropriate properties (e.g. P98_broght_into_life <child>, P96_by_mother <mother>, P97_from_father <father>).

- AgRelOn relations include hasSpouse (between the husband and wife), hasSibling (between the two brothers), hasChild (from mother/father to child), hasNieceNephew (from the brother to the child). AgRelOn relations abstract over gender (which drastically reduces the number of required relations) and are always bidirectional (each relation has inverse or is symmetric).

Person Names

Names are some of the most important personal characteristics that allow identification, especially because a lot of the other features in USHMM Persons are sparsely populated.

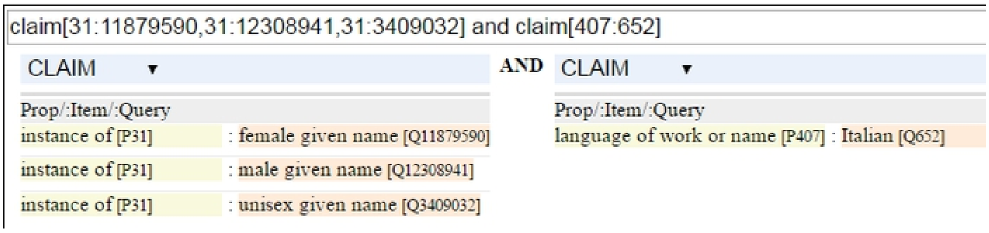
We are currently investigating what databases of names and name variants we can use. Useful information and considerations include:

- First names (given names), last names (families or surnames), middle names (and rules for forming patronymics in various cultures)
- Alternative spellings in a variety of scripts and languages, *e.g.* transliteration from Hebrew script into Latin
- Normalization rules (*e.g.* Unicode normalization for German umlauts, *e.g.* Niemoeller=Niemöller), Approximate matching (*e.g.*: Malvine=Malvina=Malvin)
- Name variants (derivations) in various cultures, *e.g.*: Giulia (Italian) is similar to Ульяна=Uljana (Russian/Ukrainian) and is similar to Юлия=Julija=Yulya=Yuliya (Russian/Slavic languages)
- Gender (for first names)
- Gender-specific morphological rules (for middle and last names). *E.g.,* the wife of a Slavic man Ivanov (Иванов) would be Ivanova (Иванова); whereas in Icelandic the suffixes *son and *dottir are often used
- Culture- and period-specific probabilities for adopting a different name at marriage (*e.g.* keeping maiden name, adopting married name, or adopting a doubled name)
- Diminutive variants; *e.g.* Yakov is a male first name, which for a boy or adult, in Russia, could be written Yashka (ru-Latn) or Яшка (ru-Cyrl)

We are currently investigating the applicability of various data sources, e.g.:

- Ancestry.com's name dictionaries for various languages. Ancestry's tool is used for keying part of the USHMM Person info.

- Yad Vashem's comprehensive database of Jewish names.
- Wikidata, which has relevant classes, e.g. “given name” and subclasses “male given name”, “female given name” and “unisex given name”. E.g. the following WDQ query[19] returns a list of 308 Italian given names, including corresponding names in other cultures. The cryptic query “claim[31:11879590,31:12308941,31:3409032] and claim[407:652]” can be decoded at the WDQ tool[20] and means the following:



claim[31:11879590,31:12308941,31:3409032] and claim[407:652]

CLAIM ▾		AND	CLAIM ▾
Prop/:Item/:Query			Prop/:Item/:Query
instance of [P31]	: female given name [Q11879590]		language of work or name [P407] : Italian [Q652]
instance of [P31]	: male given name [Q12308941]		
instance of [P31]	: unisex given name [Q3409032]		

Picture 6: Searching for Names on Wikidata. Courtesy WDQ tool, 2016

The same information can also be obtained from other sources, i.e. the Wikidata project “Names” [21]coordinates work on gathering and describing names, and has interesting info & resources. We haven't yet assessed how many variants are available.

Finally, we could use various Wikipedia categories and lists to extract raw names, e.g.: https://en.wikipedia.org/wiki/Category:Jewish_given_names ; https://en.wikipedia.org/wiki/Category:Jewish_surnames. Unfortunately these names, e.g. Abendana,[22] have no Inter-Language Links, i.e., the respective Wikidata item has only an English label.

Research Data Infrastructures

WP13 (Data Infrastructures) has the goal to “productise” useful results from WP14 by turning them into optimized, higher-capacity, managed web services that can be used by researchers and IT people in other WPs for various common tasks.

It also builds up domain knowledge bases from institutional databases by using deduplication, semantic data integration, and semantic text analysis. A few examples of specialised Natural Language Processing (NLP) that can add a lot of factual info.

[19] http://tools.wmflabs.org/autolist/autolist1.html?lang=it&q=claim%5B31%3A11879590%2C31%3A12308941%2C31%3A3409032%5D%20and%20claim%5B407%3A652%5D

[20] http://wdq.wmflabs.org/wdq/

[21] https://www.wikidata.org/wiki/Wikidata:WikiProject_Names

[22] https://en.wikipedia.org/wiki/Abendana

Very few archives provide descriptions of historical agents in the EAC format (see next section). More of them provide some structured biographical info in EAD, but it is quite sparse. However, the <bioghist> EAD field often includes a richer biography. E.g., the finding aid RG-31.094_01_fnd[23] includes the following biography:

> Solomon Benedictovich Telingater was a noted graphic artist. He was born on May 12, 1903 in Tblisi, Georgia to parents Benedict Rafaelovich Telingater and Sara Itzkovna Telingater (nee Mintzer). His family moved to Baku in 1910, where Solomon Benedictovich first learned to draw. In 1920 he finished art school in Azerbaijan, and later worked from 1921 to 1925 for the newspapers "Young Worker" and "Burden" in Baku. In 1926, Solomon moved to Moscow and began working in the publishing industry. From 1927 on, his artwork began showing up in international exhibits. His graphic artistry showed up in exhibits from New York to Paris. During World War II, Solomon created artwork to support the Soviet war effort and received the Order of the Red Star in recognition of this work. After the war, and for the remainder of his life, he continued producing artwork and participating in international exhibits. He received further awards and medals in recognition of his life work, and many of his works have been reproduced and published in books. Solomon Benedictovich Telingater died on October 1, 1969 in Moscow of a stroke.

We have underlined text that can feasibly be parsed to provide factual info. We can identify the following kinds of info:

- Full events, including type of activity, date and place. The birth and death events are especially important.
- Familial relations, especially mother and father. Additional info about the parents can also be gleaned, e.g. gender (father is male, mother is female), mother's maiden name (*nee*).
- Profession/occupation, awards, etc.
- Dates and places for which we may not be able to determine event type, but nevertheless outline the "life trajectory" of the person.

The biographical facts complement nicely (and compare favorably to) the access point provided explicitly for this person:

- Corporate Name: Soviet Union. Raboche-Krestʹi͡anskai͡a Krasnai͡a Armii͡a
- Geographic Name: Moscow (Russia)
- Personal Name: Telingater, Solomon Benedictovich, 1903-1969.
- Topical Term: Jewish artists--Biography.

[23] http://collections.ushmm.org/findingaids/RG-31.094_01_fnd_en.pdf

- Topical Term: World War, 1939-1945--Participation, Jewish--Ukraine.
- Topical Term: Jews--Ukraine--Memoirs.

Many of the lists (sources) in the USHMM Persons database are associated with a historic event. Parsing the factual info about this event can provide very useful "default" info for the persons in the list, as described previously.

Finally, we could attempt event recognition from Oral History interviews. However this task will be especially difficult, since interview transcripts use a very informal style, with interjections, interlocutions, repetitions, misspellings, etc. E.g. consider:

- Transcript RG-50.005.0047_trs:[24] "*I was born in a town which is named Pyotr Trebnazi (sp) in Poland. Its about 49 kilometers from Ladj*": the actual place is Piotrkow Trybunalski near Lodz.
- Transcript RG-50.488.0016_trs:[25] "*Q: And do you know anything about this child hidden in Bełżec? A: It was at my aunt's, my aunt was hiding. Q: Whom? A: Yes, it was this Helman's sister, Salka with her child. She was hiding them*". It's hard even for a person to understand the family relations here, let alone a machine.

WP13 also will address issues of longevity, permanence (data preservation), reproducibility of research, citability and reusability of research data. Finally, it will define and implement the EHRI Architecture, addressing issues like privacy/security, reusability for a variety of Virtual Research Environments, etc.

Person Deduplication

The USHMM Holocaust Survivors and Victims Database[26] contains over 3.200.000 person records (as of June 2015) gathered over time from heterogeneous sources. Some of the persons in the database are represented by more than one such record. The records come from different sources - lists of arrests, lists of convoys, etc. Different records may contain various information - some may be more complete than others and even if a bunch of records contain information about the same person, they are not linked to each other. For example, consider the following two records: Zoltan Grun (1)[27] and Zoltan Grun (2)[28]

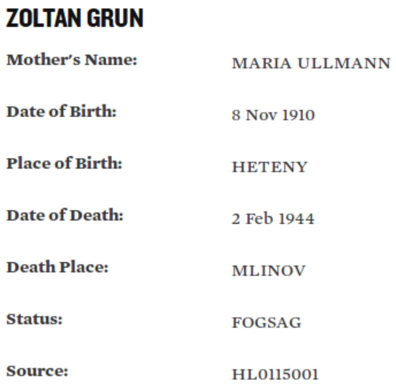
[24] http://collections.ushmm.org/oh_findingaids/RG-50.005.0047_trs_en.pdf

[25] http://collections.ushmm.org/oh_findingaids/RG-50.488.0016_trs_en.pdf

[26] https://www.ushmm.org/remember/the-holocaust-survivors-and-victims-resource-center/holocaust-survivors-and-victims-database

[27] https://www.ushmm.org/online/hsv/person_view.php?PersonId=3428503

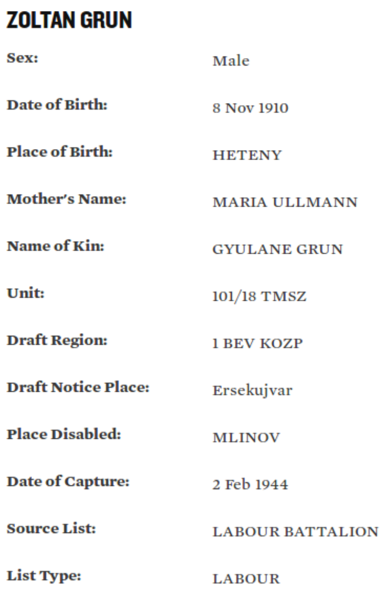
[28] https://www.ushmm.org/online/hsv/person_view.php?PersonId=5843067



ZOLTAN GRUN

Mother's Name:	MARIA ULLMANN
Date of Birth:	8 Nov 1910
Place of Birth:	HETENY
Date of Death:	2 Feb 1944
Death Place:	MLINOV
Status:	FOGSAG
Source:	HL0115001

Picture 7: Zoltan Grun (1)



ZOLTAN GRUN

Sex:	Male
Date of Birth:	8 Nov 1910
Place of Birth:	HETENY
Mother's Name:	MARIA ULLMANN
Name of Kin:	GYULANE GRUN
Unit:	101/18 TMSZ
Draft Region:	1 BEV KOZP
Draft Notice Place:	Ersekujvar
Place Disabled:	MLINOV
Date of Capture:	2 Feb 1944
Source List:	LABOUR BATTALION
List Type:	LABOUR

Picture 8: Zoltan Grun (2)

Since the exploration of person data is one of the most common tasks in the Holocaust research, we consider it important to find an automatized way to link these records and allow for better search and discovery of data related to the survivors and victims of the Holocaust. We call this task *record linking* or also *record deduplication*. The final goal of this task is to have a single record representing each person that contains the references to all the separate records related to him available from the different sources.

We solve this problem by a combination of statistical and rule-based models which could be roughly split into the following tasks:

- **Record normalization**: we transform the properties to their normalized values.
- **Classification**: we classify each person record against a number of similar candidates to find out whether they are indeed duplicates or not.
- **Clustering**: we use the result of the classification step as a distance measure between the records and cluster them in groups. Each group of records represent a single person.

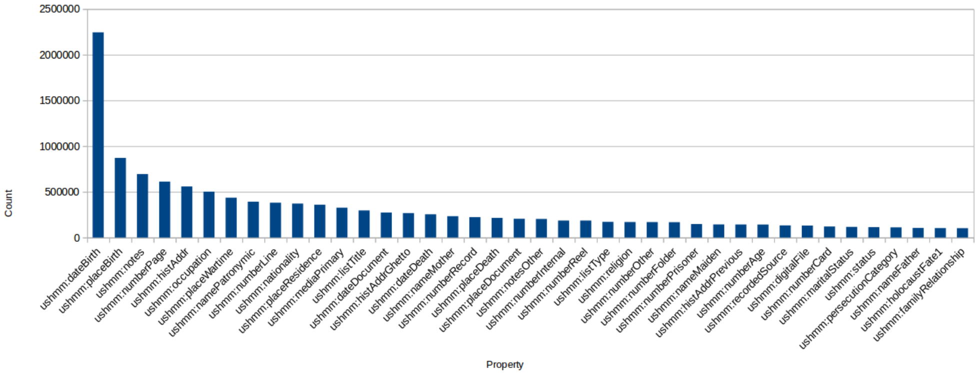
These tasks are explained in detail further on.

Data

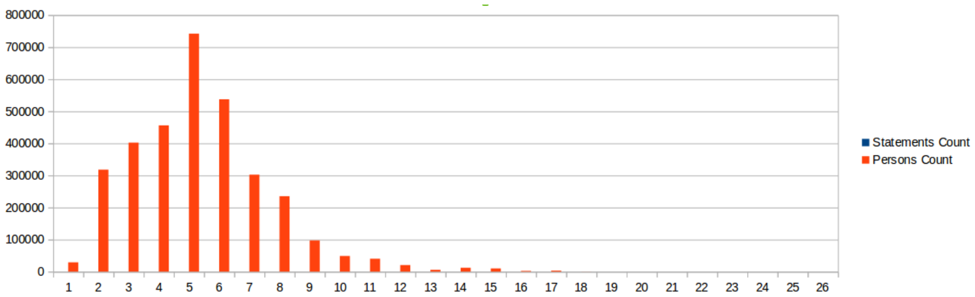
Each person record has up to 375 properties (name, date of birth, date of death, mother's name, etc.) however most of them are datatype properties i.e. literals and can contain various transliterations, languages, list of values etc. which hamper the matching between records.

- *E.g., place_of_birth*: "Mukacevo (Czechoslovakia), Czechoslovakia" vs. "Mukačevo" - being the same location
- *name*: "Schmil Zelinsky" vs. "Schmul Zelinsky" - being the same person

In addition to the variation of literals, the properties are sparsely distributed which could be observed in the figures below. The first 3 most popular properties are dateBirth, placeBirth, notes



Picture 9: Distribution of properties appearing over 10 000 times in the database



Picture 10: Distribution of the number of statements per person record

There are only 40 properties appearing more than 10 000 times in the database of over 3M records (Picture 9) and 96% of the records have between 2 and 9 properties only (Picture 10). Given these figures, we could rely on a very limited number of properties which could be used for making a decision whether two records are duplicates or not.

Data Cleanup and Normalization

Before using the datatype properties as features for finding similar records we need to normalize them. The normalized values are recorded back as new properties in the database. The procedures include the following aspects.

Name normalization

All non-Latin characters, accented and special characters occurring in names are converted to ASCII in order to improve the probability of matching.

- Convert the name to lowercase.
- Remove all digits (0 - 9) and dashes ("–", "–", "–") from the name.
- Transliterate all symbols to Latin.
- Convert the name into Unicode NFKD normal form.
- Remove all combining characters, e.g. apostrophes ("'", "'", "`"), spacing modifier letters, combining diacritical marks, combining diacritical marks supplement, combining diacritical marks for symbols, combining half marks.
- Replace punctuation characters (One of !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~) with spaces.
- Replace sequences of spaces with a single space.
- Remove any character different from space or lower case Latin letter.
- Remove all preceding or trailing white spaces from the name.
- Convert the name to title case.

Gender Prediction

The gender of a person is an important characteristic when judging whether two records correspond to the same person however it is often missing. We apply two different procedures for predicting the person gender and fill in the missing values in the database.

We train a statistical model for automatic labelling of the person gender based on the features: name suffixes with various length (1, 2 and 3); first name; concatenation of first and last name; nationality.

In addition to the statistical approach we apply also the following rules for gender induction:

- If there is another person with the same whole name and known gender, assign the same gender to the person with unknown gender.
- If there is another person with the same first name and known gender, assign the same gender to the person with unknown gender.
- If there are persons with different genders and holding the same name, do not assign a gender for the records having that name and unknown gender.

Linking USHMM Places to Geonames

Properties describing the place of birth, place of death of place of arrest and other locations are also important for the deduplication task and since they are also available as datatype properties only, we normalize them by resolving them to GeoNames locations as described in Geo-Referencing Service and Geographic Challenges. The GeoNames URIs are also added to the database and used to form the attributes for training our statistical model.

Gold Standard

The gold corpus for the classification tasks consists of 1508 pairs of person records and a label for each pair. The labels are as follows:

1. **-2** = **no** - there is a statement in the records that proves that both persons are not duplicates. *E.g.* one has been liberated in Haifa, the other one has left with a ship to Australia; the dates of birth differ too much etc.
2. **-1** = **maybe no** - there is a statement in the records that suggest that both persons are not duplicates however it is still not a certain prove.
3. **0** = **no info** - there is not enough information provided in the properties to be able to judge whether the two persons are duplicates or not.
4. **1** = **maybe yes** - there is a statement in the records that suggest that both persons are duplicates however this is still not a certain prove.
5. **2** = **yes** - there is a statement in the records that proves that the two persons are certainly duplicates.

The first 1000 records were extracted following the procedure:

- Sort all person records in descending order by count of the statements in which the person is the subject.
- Select the top 1000 persons
- Use IdRF v4.2.0 to find candidates for each person: iterative search for candidates with different rules (same first name and same last name and similar birth date and similar place of birth; same first name and same last name and similar birth date; etc).
- For each person select the candidate with the most statements.

Then additional 500 records were extracted following the procedure:

1) Select persons with first name, last name, date of birth and place of birth (as most important properties)
2) Group them by place and date of birth.
3) Select the first 500 person pairs from the groups.

The final 9 examples are positive examples which are selected *ad hoc*.

Method

Statistical model for automatic labelling of duplicated records

In our task, one classification instance is the vectorized representation of a pair of person records. This vector characterizes the similarity between two records. The algorithms takes a classification instance as input and outputs a label whether both records are referring to the same person or not. We simplify classes presented in the Gold Standard section to NO (comprising -2, -1), UNCERTAIN and YES (comprising 1 and 2). We use a multithreaded sigmoid perceptron classifier and train it on the following feature space:

- Lexical similarities of names (person first and last name, mother name): Jaro-Winkler Similarity; Levenshtein Similarity; Beider Morse Phonetic Codes Matching
- Birth dates similarity
- Birth place similarity - for finding similarity in locations we use several features: Lexical similarity; Linked Geonames instances match or not; Hierarchical structure of Geonames in order to predict the country where data is missing
- Genders Match (Male, Female)
- Person Types Match (victim, survivor, rescuer, liberator, witness, relative, veteran)
- Occupations Match
- Nationalities Match

When a new pair of records in submitted to the classifier, it outputs a probability value for this pair to fall into class YES, NO or UNCERTAIN. We use the probability to be labeled with the positive class YES as a distance measure to cluster the records in the next step.

Clustering of USHMM Person Records

In our setting, each cluster represents a unique person among the set of all clusters. In one cluster there may be one or many person's records. Each clustering algorithm requires a measure of distance between two instances in order to decide whether they are close enough to be in the same cluster. We use the probability assigned by the classifier for the given pair of instances to be labeled with the positive class YES as a distance measure. The clustering algorithm is DBSCAN.

Results

Classification

The best obtained model for classifying pairs of person records achieved the following overall mean scores: 93% F1 for the positive class YES, 88% F1 for the negative class NO and 43% F1 for the class UNCERTAIN. The mean macro F1 score is 75%.

Clustering

The algorithm for clustering was first validated on the training data and then run on the full database with the following parameters.

Input parameters:

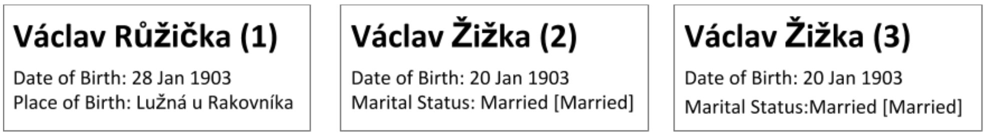
- Epsilon : 0.1
- Minimal points : 2
- Levenshtein distance : 0.15
- Reachable points : 1034623 from 3622508

Output measures

- Average points in cluster : 1.2608588509
- Total clusters : 285163
- Biggest cluster size : 1490
- Smallest cluster size : 1

Here are some examples of the clusters.

Example 1. A cluster of 3 records: The family name in the first record is different from the other two but it sounds suspiciously close and his birth date is also very close to the others.



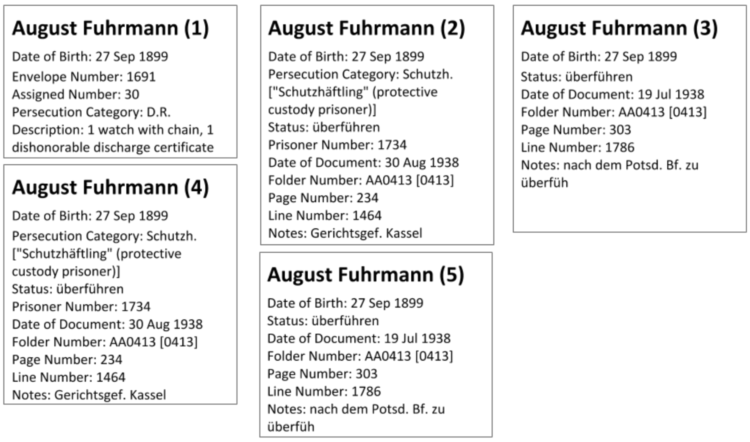
Václav Růžička (1)
Date of Birth: 28 Jan 1903
Place of Birth: Lužná u Rakovníka

Václav Žižka (2)
Date of Birth: 20 Jan 1903
Marital Status: Married [Married]

Václav Žižka (3)
Date of Birth: 20 Jan 1903
Marital Status:Married [Married]

Picture 11: https://www.ushmm.org/online/hsv/person_view.php?PersonId=6434661; https://www.ushmm.org/online/hsv/person_view.php?PersonId=6429128; https://www.ushmm.org/online/hsv/person_view.php?PersonId=6418737

Example 2. A cluster of 5 records: All records seem to refer to the same person, August Fuhrmann



August Fuhrmann (1)
Date of Birth: 27 Sep 1899
Envelope Number: 1691
Assigned Number: 30
Persecution Category: D.R.
Description: 1 watch with chain, 1 dishonorable discharge certificate

August Fuhrmann (4)
Date of Birth: 27 Sep 1899
Persecution Category: Schutzh. ["Schutzhäftling" (protective custody prisoner)]
Status: überführen
Prisoner Number: 1734
Date of Document: 30 Aug 1938
Folder Number: AA0413 [0413]
Page Number: 234
Line Number: 1464
Notes: Gerichtsgef. Kassel

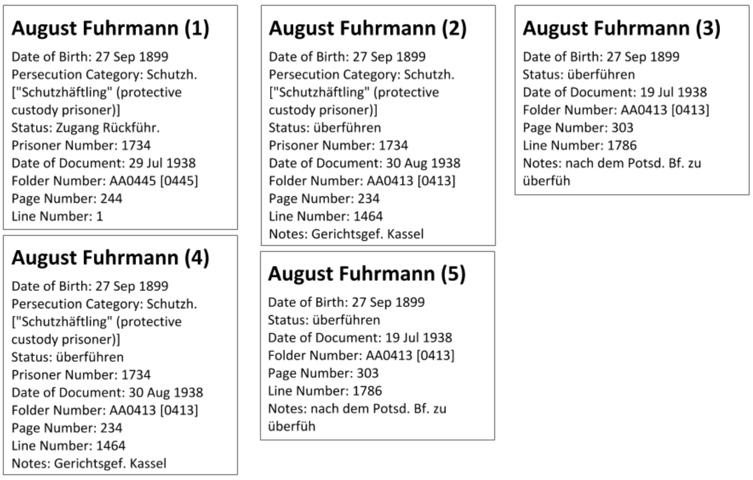
August Fuhrmann (2)
Date of Birth: 27 Sep 1899
Persecution Category: Schutzh. ["Schutzhäftling" (protective custody prisoner)]
Status: überführen
Prisoner Number: 1734
Date of Document: 30 Aug 1938
Folder Number: AA0413 [0413]
Page Number: 234
Line Number: 1464
Notes: Gerichtsgef. Kassel

August Fuhrmann (5)
Date of Birth: 27 Sep 1899
Status: überführen
Date of Document: 19 Jul 1938
Folder Number: AA0413 [0413]
Page Number: 303
Line Number: 1786
Notes: nach dem Potsd. Bf. zu überfüh

August Fuhrmann (3)
Date of Birth: 27 Sep 1899
Status: überführen
Date of Document: 19 Jul 1938
Folder Number: AA0413 [0413]
Page Number: 303
Line Number: 1786
Notes: nach dem Potsd. Bf. zu überfüh

Picture 12: https://www.ushmm.org/online/hsv/person_view.php?PersonId=2996015; https://www.ushmm.org/online/hsv/person_view.php?PersonId=4430175; https://www.ushmm.org/online/hsv/person_view.php?PersonId=4433342; https://www.ushmm.org/online/hsv/person_view.php?PersonId=4433343; https://www.ushmm.org/online/hsv/person_view.php?PersonId=4433344

Example 3. A cluster of 5 records: Here we see a separate cluster which seems to contain records referring again to the same August Fuhrmann and needs to be merged with the cluster in Example 2 (Picture 12).



August Fuhrmann (1)
Date of Birth: 27 Sep 1899
Persecution Category: Schutzh. ["Schutzhäftling" (protective custody prisoner)]
Status: Zugang Rückführ.
Prisoner Number: 1734
Date of Document: 29 Jul 1938
Folder Number: AA0445 [0445]
Page Number: 244
Line Number: 1

August Fuhrmann (2)
Date of Birth: 27 Sep 1899
Persecution Category: Schutzh. ["Schutzhäftling" (protective custody prisoner)]
Status: überführen
Prisoner Number: 1734
Date of Document: 30 Aug 1938
Folder Number: AA0413 [0413]
Page Number: 234
Line Number: 1464
Notes: Gerichtsgef. Kassel

August Fuhrmann (3)
Date of Birth: 27 Sep 1899
Status: überführen
Date of Document: 19 Jul 1938
Folder Number: AA0413 [0413]
Page Number: 303
Line Number: 1786
Notes: nach dem Potsd. Bf. zu überfüh

August Fuhrmann (4)
Date of Birth: 27 Sep 1899
Persecution Category: Schutzh. ["Schutzhäftling" (protective custody prisoner)]
Status: überführen
Prisoner Number: 1734
Date of Document: 30 Aug 1938
Folder Number: AA0413 [0413]
Page Number: 234
Line Number: 1464
Notes: Gerichtsgef. Kassel

August Fuhrmann (5)
Date of Birth: 27 Sep 1899
Status: überführen
Date of Document: 19 Jul 1938
Folder Number: AA0413 [0413]
Page Number: 303
Line Number: 1786
Notes: nach dem Potsd. Bf. zu überfüh

Picture 13: https://www.ushmm.org/online/hsv/person_view.php?PersonId=4433345; https://www.ushmm.org/online/hsv/person_view.php?PersonId=4442252; https://www.ushmm.org/online/hsv/person_view.php?PersonId=4442559; https://www.ushmm.org/online/hsv/person_view.php?PersonId=4443021 ; https://www.ushmm.org/online/hsv/person_view.php?PersonId=4443328

Thorough evaluation of the clusters and error analysis is still to be conducted and if necessary adjustments to the clustering procedures can be done.

We showed that the task of person deduplication is feasible and could be performed with quite a high accuracy. With the help of such a procedure, human experts could interlink the records of over 250 000 people who were involved in the Holocaust, thereby improving the retrieval of documents regarding concrete personalities.

EAD Archival Descriptions

So far EHRI has aggregated 201,214 archival descriptions[29] from 463 institutions,[30] representing

18,308 archival units. Also, EHRI has developed 57 country reports[31] outlining the brief history of the country's involvement in WW2 and the Holocaust, and the archival situation (main archival materials and institutions) in the country. EHRI has aggregated detailed info about 1,847 archival institutions.[32] This information is developed following established archival content standards:

- ISAD(G) for describing archival units

[29] https://portal.ehri-project.eu/units
[30] https://portal.ehri-project.eu/institutions?data=yes
[31] https://portal.ehri-project.eu/countries
[32] https://portal.ehri-project.eu/institutions

- ISAAR for describing the context of creation: Personalities, Organisations (including Nazi administrative districts), Events, Camps, Ghettos
- ISDIAH for describing collection holding institutions
- The respective technical XML schemas and tag libraries that formalize these descriptions are:
- EAD (Encoded Archival Description) for archival units
- EAC/CPF (Encoded Archival Context: Corporate bodies, Persons and Families) for agents
- EAG (Encoded Archival Guide) for institutions

A lot of this work was performed during EHRI-1, and the work continues in EHRI-2. A lot of manual effort was spent in EHRI-1 to aggregate the descriptions, because there was no tooling or procedures to make the process smoother and more repeatable.

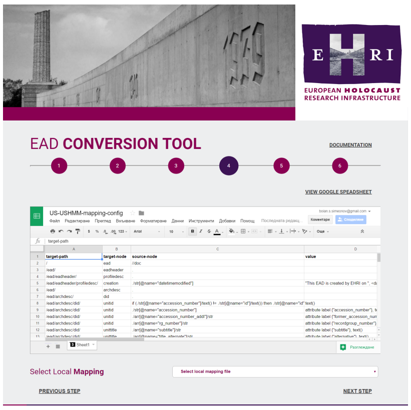
In EHRI-2, the goal of WP10 is to develop tools and procedures to make EAD aggregation more sustainable, to enable incremental update (synchronization), and to provide some self-service functionality so archival institutions can initiate the process and validate the results themselves. WP10 is structured as follows.

EAD Conversion

T10.1 developed tooling to convert archival descriptions from various formats to standard EAD XML. We selected XQuery for implementing the transformations, and Java & JavaScript for developing an archivist-friendly front-end. We use the BaseX XQuery implementation, since it has good performance, and an big library of useful extensions, e.g. writing multiple files out of one XML input.

The tool addresses tabular data and XML data (e.g. from Faust archivist databases, SOLR indexes, etc). It provides fixes for various commonly occurring problems per institution, and some special processing tasks (e.g. construction of hierarchies, links to digital material, etc). It has a step-by-step wizard interface and supports a sophisticated conversion mechanism that is driven by mappings expressed in Google sheets. Extensive documentation[33] and online tutorials are available 8.. The tool is open sourced and is available on Github (https://github.com/EHRI/ehri-conversion-tools)

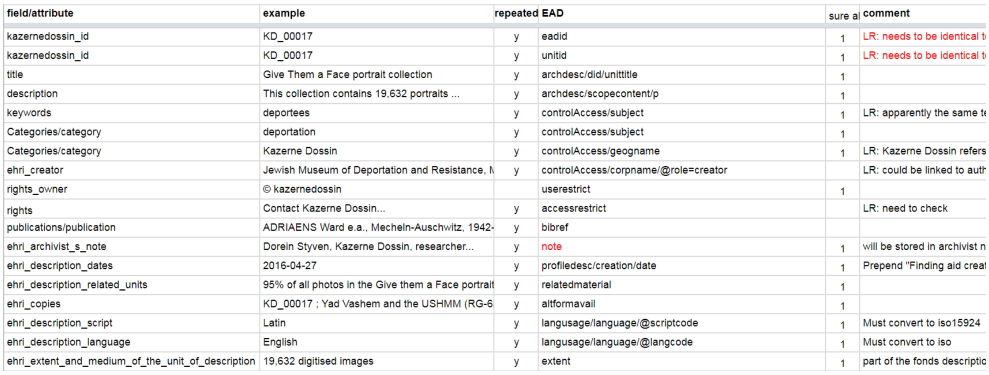
[33] https://docs.google.com/document/d/1qO2L2N_UeTeTafZxa7nKWDyZvYCZnqHi8RPhCsHMtmg/edit



	A	B	C	D
1	target-path	target-node	source-node	value
2	/	ead	//doc	
3	/ead/	eadheader	.	
4	/ead/eadheader/	profiledesc	.	
5	/ead/eadheader/profiledesc/	creation	./str[@name="datetimemodified"]	"This EAD is created by EHRI on ", <d
6	/ead/	archdesc	.	
7	/ead/archdesc/	did	.	
8	/ead/archdesc/did/	unitid	if (./str[@name="accession_number"]/text() != ./str[@name="id"]/text()) then ./str[@name="id"	text()
9	/ead/archdesc/did/	unitid	./str[@name="accession_number"]	attribute label {"accession_number"}, te
10	/ead/archdesc/did/	unitid	./arr[@name="accession_number_addl"]/str	attribute label {"former_accession_num
11	/ead/archdesc/did/	unitid	./arr[@name="rg_number"]/str	attribute label {"recordgroup_number"},
12	/ead/archdesc/did/	unittitle	./arr[@name="subtitle"]/str	attribute label {"subtitle"}, text()
13	/ead/archdesc/did/	unittitle	./arr[@name="title_alternate"]/str	attribute label {"alternative"}, text()

Picture 14: EHRI EAD Conversion Tool. Courtesy Ontotext Corp

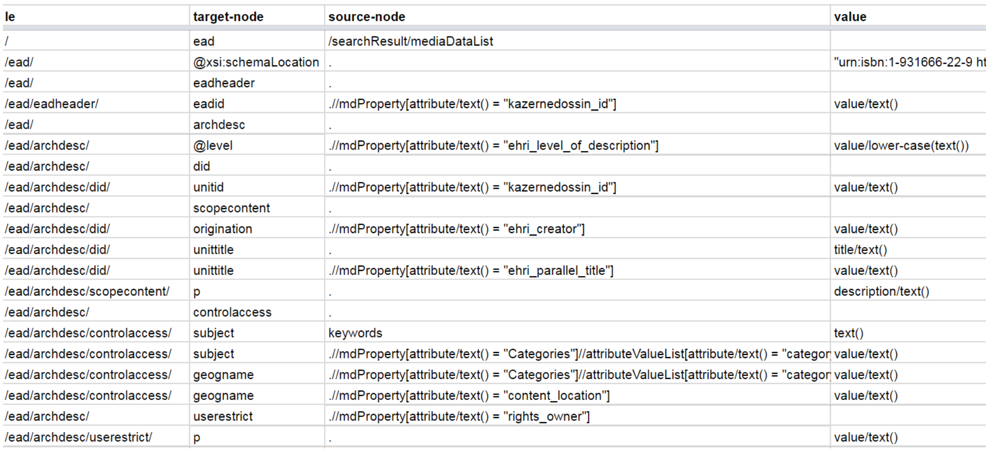
For each institution we first develop a conceptual mapping from the institution data to EAD.



field/attribute	example	repeated	EAD	sure al	comment
kazernedossin_id	KD_00017	y	eadid	1	LR: needs to be identical t
kazernedossin_id	KD_00017	y	unitid	1	LR: needs to be identical t
title	Give Them a Face portrait collection	y	archdesc/did/unittitle	1	
description	This collection contains 19,632 portraits ...	y	archdesc/scopecontent/p	1	
keywords	deportees	y	controlAccess/subject	1	LR: apparently the same te
Categories/category	deportation	y	controlAccess/subject	1	
Categories/category	Kazerne Dossin	y	controlAccess/geogname	1	LR: Kazerne Dossin refers
ehri_creator	Jewish Museum of Deportation and Resistance, M	y	controlAccess/corpname/@role=creator		LR: could be linked to auth
rights_owner	© kazernedossin		userestrict	1	
rights	Contact Kazerne Dossin...	y	accessrestrict		LR: need to check
publications/publication	ADRIAENS Ward e.a., Mecheln-Auschwitz, 1942-	y	bibref		
ehri_archivist_s_note	Dorein Styven, Kazerne Dossin, researcher...	y	note	1	will be stored in archivist n
ehri_description_dates	2016-04-27	y	profiledesc/creation/date	1	Prepend "Finding aid creat
ehri_description_related_units	95% of all photos in the Give them a Face portrait	y	relatedmaterial	1	
ehri_copies	KD_00017 ; Yad Vashem and the USHMM (RG-6	y	altformavail	1	
ehri_description_script	Latin	y	langusage/language/@scriptcode	1	Must convert to iso15924
ehri_description_language	English	y	langusage/language/@langcode	1	Must convert to iso
ehri_extent_and_medium_of_the_unit_of_description	19,632 digitised images	y	extent	1	part of the fonds descriptio

Picture 15: Conceptual Mapping of Kazerne Dossin data to EAD

We then develop a physical mapping that implements the conceptual mapping using Xpaths.



le	target-node	source-node	value
/	ead	/searchResult/mediaDataList	
/ead/	@xsi:schemaLocation	.	"urn:isbn:1-931666-22-9 ht
/ead/	eadheader	.	
/ead/eadheader/	eadid	.//mdProperty[attribute/text() = "kazernedossin_id"]	value/text()
/ead/	archdesc	.	
/ead/archdesc/	@level	.//mdProperty[attribute/text() = "ehri_level_of_description"]	value/lower-case(text())
/ead/archdesc/	did	.	
/ead/archdesc/did/	unitid	.//mdProperty[attribute/text() = "kazernedossin_id"]	value/text()
/ead/archdesc/	scopecontent	.	
/ead/archdesc/did/	origination	.//mdProperty[attribute/text() = "ehri_creator"]	value/text()
/ead/archdesc/did/	unittitle	.	title/text()
/ead/archdesc/did/	unittitle	.//mdProperty[attribute/text() = "ehri_parallel_title"]	value/text()
/ead/archdesc/scopecontent/	p	.	description/text()
/ead/archdesc/	controlaccess	.	
/ead/archdesc/controlaccess/	subject	keywords	text()
/ead/archdesc/controlaccess/	subject	.//mdProperty[attribute/text() = "Categories"]//attributeValueList[attribute/text() = "categor	value/text()
/ead/archdesc/controlaccess/	geogname	.//mdProperty[attribute/text() = "Categories"]//attributeValueList[attribute/text() = "categor	value/text()
/ead/archdesc/controlaccess/	geogname	.//mdProperty[attribute/text() = "content_location"]	value/text()
/ead/archdesc/	userestrict	.//mdProperty[attribute/text() = "rights_owner"]	
/ead/archdesc/userestrict/	p	.	value/text()

Picture 16: Physical Mapping of Kazerne Dossin data to EAD

This mapping then can be selected by the user in the EAD conversion tool, which uses it online from Google Sheets. This ensures that the mapping is always up to date. The tool also generates transformation scripts that can be executed in batch mode, e.g. as a scheduled task.

We are even considering for conversion some textual descriptions that use consistent field names, e.g. 1994.62_01_fnd.[34]

EAD Validation

The EAD tool also provides EAD validation and preview as HTML. We perform validation at 3 points:

- As final step of the conversion, so the archivist can check his work for errors
- Before publishing (transporting) a set of EAD to the EHRI portal (see the next section)
- Before ingesting EAD to the EHRI portal

Validation rules are provisioned using TEI ODD[35] files, generate RelaxNG and Schematron (see below) and provide links to EAD taglib documentation for better diagnosing of the error: https://github.com/EHRI/data-validations. We validate EAD against two different set of rules:

- The official EAD2002 schema, plus a few EHRI-specific rules (e.g. each documentary unit should have an id)
- An internal EHRI EAD schema. It's almost the same, but has some EHRI specific adaptations (e.g. paragraphs on the EHRI portal are not formatted with

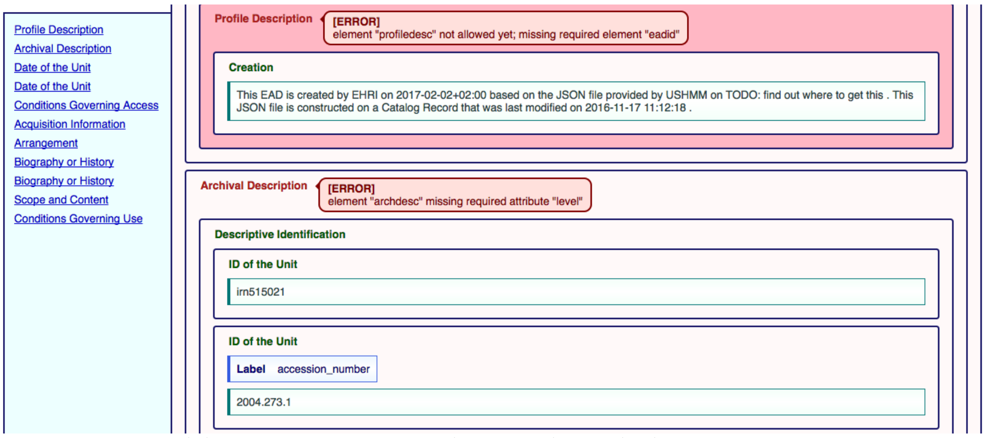
[34] http://collections.ushmm.org/findingaids/1994.62_01_fnd_en.pdf

[35] TEI: Getting Started with ODDs. http://www.tei-c.org/Guidelines/Customization/odds.xml

This allows a variety of data flows to be implemented in different combinations: using the conversion tool, using the publishing tool, or direct deposit to the EHRI ingest directory.

- We validate EAD using both RelaxNG XML schemas and Schematron rule checking. RelaxNG is a better way to express XML schemas than XSD, and RelaxNG Compact is an intuitive and compact notation (e.g. see https://github.com/VladimirAlexiev/rnc). We made RelaxNG from the official EAD2002 schema (whereas EAD3 is developed completely in RelaxNG. This checks the structural integrity of EAD files.
- We also use Schematron[36] rules for additional checks. These are used for cross-field validation, conditional checks, etc.
- We extended the popular schema jing-trang validator to report schema errors in the Schematron Validation Report Language (SVRL), which is a simple XML format to describe the source of error using XPath. See https://github.com/EHRI/jing-trang and in particular pull request 1.[37] This way the validator deals with one kind of error, no matter whether it's a schema or Schematron rule violation.

The EAD validator integrates the error reports into the HTML preview, making it more suitable for a non-technical audience. First we generate an index *<timestamp>/html/index.html* showing all produced EAD files and the number of errors in each. From this index one can access the individual preview+errors. For example, the next figure shows that the "Profile Description" element is not allowed, because there is a missing "eadid" element. In order to correct the error, one must add a "eadid" element. Depending on the data flow, one can correct them in the input file, the mapping configuration, or the XML file.



Profile Description

[ERROR]
element "profiledesc" not allowed yet; missing required element "eadid"

Creation

This EAD is created by EHRI on 2017-02-02+02:00 based on the JSON file provided by USHMM on TODO: find out where to get this . This JSON file is constructed on a Catalog Record that was last modified on 2016-11-17 11:12:18 .

Archival Description

[ERROR]
element "archdesc" missing required attribute "level"

Descriptive Identification

ID of the Unit

irn515021

ID of the Unit

Label accession_number

2004.273.1

Picture 17: EAD Validation. HTML preview and integrated error display

[36] https://en.wikipedia.org/wiki/Schematron

[37] https://github.com/EHRI/jing-trang/pull/1

EAD Publishing/Transport

T10.2 has developed an EAD Publishing (transport) tool based on OAI ResourceSync.[38] OAI ResourceSync can be used to transport metadata, allowing discovery of new and updated metadata files. It is easier to deploy than OAI PMH[39] because it does not require a server, but institutions that already have an OAI PMH server can also be accommodated by EHRI. It consists of two parts:

- https://github.com/EHRI/resyto implements a GUI that allows the user to select files, applies updating (synchronization) logic and produces ResourceSync manifests. It implements the source part of the ResourceSync Framework
- https://github.com/EHRI/resydes visits preconfigured ResourceSync endpoints, gets the manifests, checks for new or updated files since the last run, and transports them to the EHRI server. It implements the destination part of the ResourceSync Framework

EAD Ingest

T10.3 will improve the existing ingest of EADs to the EHRI database. It will enable:

- updating (synchronization)
- co-referencing of textual Access Points to proper Authority references (see next section)
- preservation of links created manually in the EHRI portal
- exposing of links to finding aids and digital materials on the EHRI portal.

Authorities and Standards

WP11 involves the following tasks:

- Studying user needs (both researchers and archivists) to define improvements needed on the EHRI portal, especially regarding search.
- Selecting, applying and extending Standards and Ontologies.
- Establishing Permanent URLs to be used by the project. We need to reconsider the design of the EHRI namespace, in order to enable semantic resolution and content negotiation (the same URL serves different content depending on the content type requested) and guarantee long-term permanence.

[38] http://www.openarchives.org/rs

[39] https://www.openarchives.org/pmh/

- Building and interconnecting Authorities and Vocabularies

Regarding standards, INRIA has integrated the EAD XML schema, tag libraries (detailed descriptions and guidelines), and additional Schematron rules developed by EHRI into a wholistic document using the TEI ODD standard. This ensures better consistency between these artefacfts and can be used by other archival integration projects as well.

The EHRI portal already can serve EAD and EAC for archival and historic agent descriptions, and RDF SKOS for some of the authorities. We are also considering some ontologies to be able to export more of the EHRI data as Linked Open Data, e.g.:

- Schema, FOAF and CRM for persons
- CRM and AgRelOn for historic events and person relations. A dedicated Shoah Ontology was developed by CDEC earlier, but CRM is more appropriate for modelling persons, networks, life events, and places in a generic way.
- SKOSXL for Subject authorities.

Although the archival standards (EAD and EAC) are not semantic at all (e.g. there are no links to agents, only agent names), they can be enriched semantically, some of the info can be exposed as links, and they can be represented in RDF.

The rest of this section describes the Authority work in EHRI. It consolidates and enlarges the EHRI authorities to render the indexing and retrieval of information more effective. It addresses:

- Access Points in ingested EADs (normalization of Unicode, spelling, punctuation; deduplication; clustering; co-referencing to authority control),
- Subjects (deployment of a Thesaurus Management System),
- Places (co-referencing to Geonames);
- Camps and Ghettos (integrating data from Wikidata);
- Persons, Corporate Bodies.

We also hope to implement semantic (conceptual) search including hierarchical query expansion; interconnectivity of archival descriptions.

A lot of hard deduplication (co-referencing) work is ongoing to build up the EHRI authorities, including resolution of spelling differences and multilingual problems, co-referencing to global authorities where available, integrating disparate EHRI thesauri, deciding whether “candidate” concepts harvested from access points should be added to the thesaurus, etc. We use a mix of automatic and manual approaches, through the establishment of an EHRI Thesaurus Editorial Board.

EHRI-1 Thesauri

EHRI-1 has published linked data (http://data.ehri-project.eu/) that includes the following authorities:

- 880 concepts (the EHRI thesaurus)
- 1970 camps
- 1000 ghettos
- 400 people
- 260 Nazi administrative districts
- 50 events. These are major historic events that one can find on Wikipedia, e.g. “Blood for Goods”.[40]

Camps and Ghettos

EHRI-1 had collected a master list of Concentration Camps and Ghettos, but it included minimal information: name in a few languages (*e.g.* “Maly Trostinec”@de-latn) and in some cases sub-camp relations (hierarchy). Wikipedia and Wikidata include a lot more info about camps.

E.g., **Wikipedia** has the following info on Maly_Trostenets_extermination_camp.[41] References are provided, but the info is **not structured:**

- Names: Maly Trostinets, Maly Trastsianiets, Trasciane, Малы Трасцянец, Maly Tras’tsyanyets, Малый Тростенец, Maly Trostinez, Maly Trostenez, Maly Trostinec, Klein Trostenez
- Location: outskirts of Minsk
- Admin district: Reichskommissariat Ostland
- Established: 10 May 1942
- Killing grounds: Blagovshchina (Благовщина) forest, Shashkovka (Шашковка) forest
- Victim countries: predominantly Belarus (inferred, not explicitly stated). Also Austria, Germany, Czech Republic.
- Victim places of origin: predominantly Minsk. Also Berlin, Hanover, Dortmund, Münster, Düsseldorf, Cologne, Frankfurt am Main, Kassel, Stuttgart, Nuremberg, Munich, Breslau, Königsberg, Vienna, Prague, Brünn, Theresienstadt.

[40] https://en.wikipedia.org/wiki/Category:Blood_for_goods

[41] https://en.wikipedia.org/wiki/Maly_Trostenets_extermination_camp

- Known victims: Vincent Hadleŭski (Wincenty Godlewski): arrested in Minsk on December 24, 1942 and shot at Trascianiec the same day; Norbert Jokl (debated); Margarete Hilferding (in transit to the camp from Terezín); Grete Forst; Cora Berliner (most likely)

- Perpetrators (and their fate): SS Unterscharführer Heinrich Eiche (fled to Argentina after the war and all trace of him was lost); Eduard Strauch (died in Belgian prison in 1955); Rottenführer Otto Erich Drews (in 1968 the Court in Hamburg sentenced to life imprisonment); Revieroberleutnant Otto Hugo Goldapp (in 1968 the Court in Hamburg sentenced to life imprisonment); Hauptsturmführer Max Hermann Richard Krahner (in 1968 the Court in Hamburg sentenced to life imprisonment); Heinrich Seetzen (committed suicide in a British POW camp); Gerhard Maywald (settled after the war in West Germany; on August 4, 1977 sentenced to 4 years imprisonment); Jewish Sonderkommando 1005

Wikidata (Q316109)[42] knows the following **structured** info:

- Names and Wikipedia links in the following languages: Беларуская, Беларуская (тарашкевіца), Čeština, Dansk, Deutsch, Español, Français, Frysk, Italiano, עברית, Nederlands, Norsk bokmål, Polski, Português, Русский, Српски / srpski, Suomi, Svenska, Українська, 中文

- Additional aliases, e.g. Vernichtungslager Maly Trostinez, KZ Maly Trostinez, Blagowschtschina

- Country: Belarus

- Location: 53°51'3"N, 27°42'17"E

- Authority identifiers: Geonames, VIAF, Freebase

DBpedia (Maly_Trostenets_extermination_camp)[43] knows the following **structured** info:

1) links to Wikidata, Geonames, Freebase, different Wikipedias
2) coordinates
3) a few more aliases: Maly_Tras'tsyanyets, Maly_Tras'tsyanyets_camp, Maly_Tras'tsyanyets_concentration_camp, Maly_Tras'tsyanyets_extermination_camp

The fact that it is DeathPlace of the following people. This comes from the articles about these people (i.e., inverse links): dbr:Margarete_Hilferding, dbr:Grete_Forst, and dbr:Vincent_Hadleŭski.
Categories: dbc:World_War_II_sites_of_Nazi_Germany, dbc:Geography_of_Minsk,

[42] https://www.wikidata.org/wiki/Q316109

[43] http://live.dbpedia.org/page/Maly_Trostenets_extermination_camp

dbc:History_of_Belarus_(1939–1945), dbc:History_of_Minsk, dbc:Maly_Trostenets_extermination_camp, dbc:The_Holocaust_in_Belarus, dbc:World_War_II_sites_in_Belarus, dbc:Belarus_in_World_War_II.

Wikidata templates about specific camps have even more info, e.g. Template:Treblinka_extermination_cam[44]p includes detailed lists of **perpetrators, victims, resistance**, etc., all linked to Wikipedia:

- **Camp organizers**: Odilo Lotario Globocnik, Hermann Julius Höfle, Erwin Hermann, Lambert Richard, Wolfgang Thomalla, Christian Wirth
- **Commandant**: Irmfried Eberl, Franz Paul Stangl, Kurt Hubert Franz
- **Deputies**: Theodor van Eupen, Heinrich Arthur Matthes Karl Pötzinger
- Gas chamber **executioners**: Gustav Münzberger, Fritz Schmidt
- Other **officers**: Max Biala, Paul Bredow, Herbert Floss, Erich Fritz Erhard Fuchs, Lorenz Hackenholt, Hans Hingst, Josef Hirtreiter, Otto Richard Horn Kurt Küttner, Karl Emil Ludwig Willy Mätzig, Willi Mentz, August Wilhelm Miete, Max Möller, Willi Post, Albert Franz Rum, Karl Schiffer, Otto Stadie, Ernst Stengelin, Franz Suchomel
- **Guards**: Ukrainians: "Ivan the Terrible", John Demjanjuk, Feodor Fedorenko, Nikolay Yegorovich Shalayev; Others: "Trawnikis", Volksdeutsche
- Prominent **victims**: Ernst Arndt, Yitzchok Breiter, Amalia Carneri, Julian Chorążycki, Samuel Finkelstein, Artur Gold Ludwik Holcman, Janusz Korczak, Berek Lajcher, Henryka Łazowertówna, Yechiel Lerer, Yitzchak Lowy, Simon Pullman, Natan Spigel, Symche Trachter, Zygmunt Zalcwasser, Lidia Zamenhof
- **Resistance**, Survivors: Richard Glazar, Chil Rajchman, Sol Rosenberg, Kalman Taigman, Jankiel Wiernik, Samuel Willenberg, Franciszek Ząbecki
- **Nazi organizations**: General Government SS-Totenkopfverbände
- **Planning Methods, Documents, Evidence**: Operation Reinhard Höfle Telegram
- **Aftermath, Memorials**: Treblinka trials
- **Related topics**: The Holocaust Operation Reinhard, Nazi concentration camps, Extermination camp

Recognizing the value of this data and working together with the LOD community, EHRI decided in Feb 2017 to enrich and integrate data on Camps and Ghettos using Wikidata as a data integration platform. We first created or enriched Wikidata entities and properties to use in the matching process, *e.g.*:

[44] https://en.wikipedia.org/wiki/Template:Treblinka_extermination_camp

- USHMM Holocaust Encyclopedia (Q5883924).[45]
  Extended the Wikipedia article Holocaust_Encyclopedia with some numbers and website https://www.ushmm.org/learn/holocaust-encyclopedia.
- Yad Vashem Encyclopedia of the Holocaust (Q2906963).[46] Represented at Wikipedia article Encyclopedia_of_the_Holocaust[47]
- EHRI Project (Q21755493)[48]
- USHMM Holocaust Encyclopedia ID (P3724):[49] identifier of camp/ghetto to link to this encyclopedia. See property proposal.[50]
- Yad Vashem Encyclopedia of the Ghettos ID (P3735):[51] identifier of a ghetto, or a place containing a ghetto. See property proposal.[52]
- USHMM person ID (P4130):[53] identifier for a person in the UHSMM HSV, a database of Holocaust survivors and victims. See property proposal.[54]

We then used data from many places:

- The two institutional encyclopedias mentioned above.
- Querying Wikidata by a set of relevant types, e.g.:

  https://www.wikidata.org/wiki/Q328468 Nazi concentration camp;

  https://www.wikidata.org/wiki/Q152081 Concentration camp;
  https://www.wikidata.org/wiki/Q1719244 Subcamp;
  https://www.wikidata.org/wiki/Q628505 Labor camp;
  https://www.wikidata.org/wiki/Q14916829 transit camp;
  https://www.wikidata.org/wiki/Q15727645 internment camp
- Querying DBpedia by a number of Yago types, e.g.:

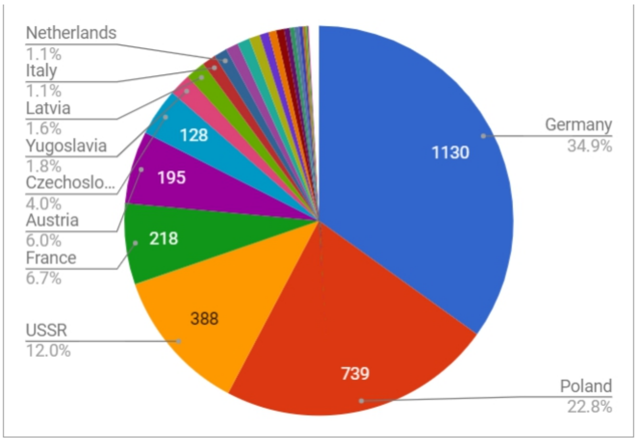
  http://dbpedia.org/class/yago/WorldWarIISitesOfNaziGermany ,
  http://dbpedia.org/class/yago/ConcentrationCamp103086183 ,
  http://dbpedia.org/class/yago/Camp102945379 ,
  http://dbpedia.org/class/yago/NaziConcentrationCamps ,
  http://dbpedia.org/class/yago/NaziConcentrationCampsInAustria ,

[45] https://www.wikidata.org/wiki/Q5883924
[46] https://www.wikidata.org/wiki/Q2906963
[47] https://en.wikipedia.org/wiki/Encyclopedia_of_the_Holocaust
[48] https://www.wikidata.org/wiki/Q21755493
[49] https://www.wikidata.org/wiki/Property:P3724
[50] https://www.wikidata.org/wiki/Wikidata:Property_proposal/USHMM_Holocaust_Encyclopedia_id
[51] https://www.wikidata.org/wiki/Property:P3735
[52] https://www.wikidata.org/wiki/Wikidata:Property_proposal/Ghetto_Encyclopedia_ID
[53] https://www.wikidata.org/wiki/Property:P4130
[54] https://www.wikidata.org/wiki/Wikidata:Property_proposal/USHMM_person_ID

http://dbpedia.org/class/yago/NaziConcentrationCampsInBelgium ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInDenmark ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInEstonia ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInFrance ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInGermany ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInItaly ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInLatvia ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInLithuania ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInNorway ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInPoland ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInTheNetherlands ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInUkraine ,
http://dbpedia.org/class/yago/NaziConcentrationCampsInYugoslavia ,
http://dbpedia.org/class/yago/NaziConcentrationCampsOnAlderney ,
http://dbpedia.org/class/yago/WorldWarIIConcentrationCamps ,
http://dbpedia.org/class/yago/WorldWarIICroatianConcentrationCampsInFormerYugoslavia

- Extracted a list of 4174 camps/ghettos from USHMM places by looking for names that include patterns like "camp|ghetto|lager|kamp|лагер|KZ" etc. We found 58 patterns that indicate a camp, e.g.: KLV camp, GTE camps?, POW camp, Prisoner of War Camps?, camp prisoner of war, camp v/pl, Labour camp, KriegsGefangenenLager, GemeinschaftsLager, K\. Lager, K\.z\. ?Lager, KZ, unbekanntes Lager, Lager unbek\., Lager unbekannt, Arbeitslager, FrauenArbeitsLager, StrafLager, Reichsbahn Lager, LAGERKÜCHE, WaldLager, Firmen?Lager, SammelLager, isprav trud lager', Isprav\. trud\. lager', Ispr\.Trud\.lager', ispr\. trudovoi lager', isprav\.trudovoi lager', Ispravitel'nyi trudovoi lager', Smertnyj Lager, Lagerei, Lageri.

We also found 20 place names that do not indicate a camp (false positives), e.g. Campagna, Campagne, Campan, Campana, Campania, Campbell, Campiniac, Campitelli, Campobasso, Campobello, Campos, Danekamp, Eichkamp, Kampen, Kampina, Kampinos, Lagerstr.



Picture 18: Count of camps by country in USHMM Places database. Courtesy Ontotext, 2018

Not all the above USHMM places represent new camps (not present in the EHRI-1 list). We performed comprehensive research (although a partial evaluation) and estimate as follows:

- 67% or 2797 places are new camps, e.g. Akmecetka (Acmecetca, Akmeketka, Акмечетка), Odessa, Ukraine/USSR;[55] Ahlem, Germany,[56] "uncovered Hannover-Ahlem (April 10, 1945) and Salzwedel (April 14, 1945), both satellite camps of Neuengamme"
- 11.34% or 473 places are duplicates, but may provide valuable altLabels, e.g. Pavlou Mela (Salonika / Makedhonía / Greece
- 15.46% or 645 are not valid camp names, e.g.: Bulgarien/unbekanntes Lager: unspecific; Fécamp: false match on name
- 3.09% or 129 are Displaced Person (DP) camps that are not Nazi camps but are relevant to the Holocaust, e.g.: Kloster Indersdorf displaced persons camp: charitable camp at a monastery; Neustadt [DP camp]: DP camp in the British zone

We also considered the CIA Nazi camps list[57] published on 7 May 1945 and titled CIA: AXIS CONCENTRATION CAMPS AND DETENTION CENTRES REPORTED AS SUCH IN EUROPE. Basic Handbook KLs (Konzentrationslager). Originally published at the CIA Library,[58] it is 187 pages long and includes info about 615 camps (and 93 cross-references) grouped by country. EHRI performed extensive cleanup and regrouping by country. See the

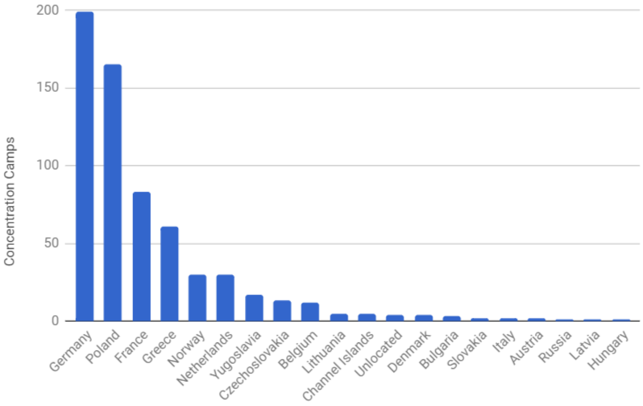
[55] Confirmed at https://ro.wikipedia.org/wiki/Lagărul_de_concentrare_Acmecetca

[56] Confirmed at http://www.tabletmag.com/scroll/194615/kissinger-on-liberating-ahlem-concentration-camp and http://faculty.jou.ufl.edu/croberts/documentary/angelofahlem/thecamp.html

[57] https://www.wikidata.org/wiki/Q39487700

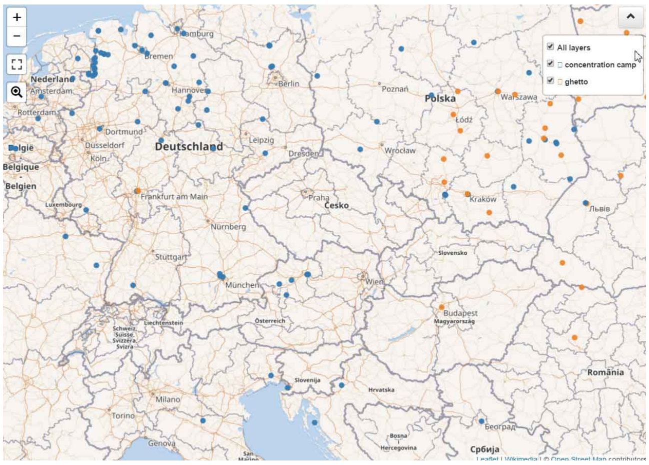
[58] https://www.cia.gov/library/readingroom/docs/GERMAN%20CONCENTRATION%20CAMPS_0001.pdf

following chart.

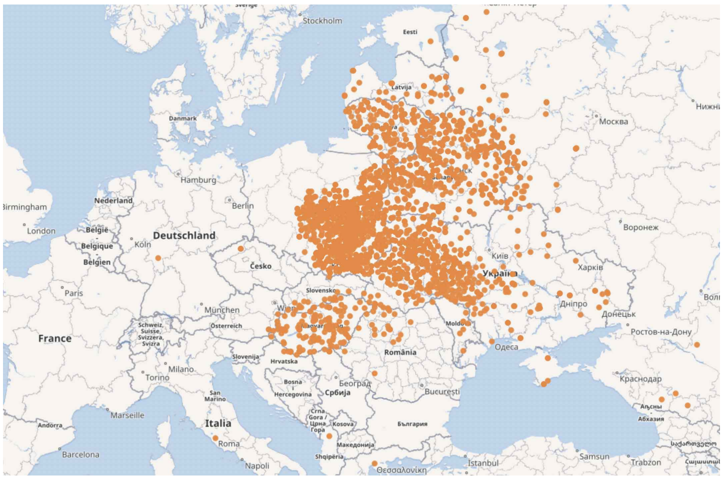


Picture 19: Count of camps by country in CIA Basic Handbook Konzentrationslager. Courtesy Ontotext, 2018

A massive matching and cross-referencing task was undertaken. As a result, we have merged several thousand Ghetto records across encyclopedias and contributed them to Wikidata, taking coordinates and associated place hierarchy in return 6.. As a result, the number of ghetto records on Wikidata has grown drastically.



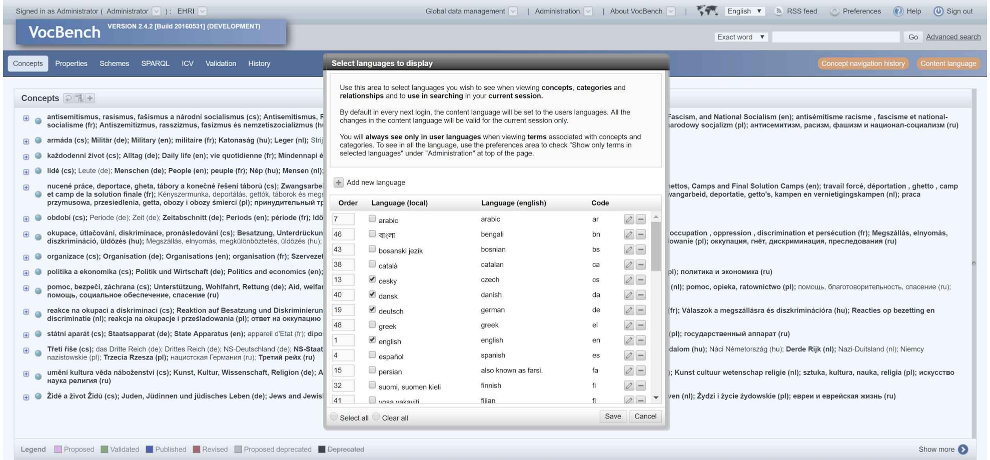
Picture 20: Ghettos (orange) and Camps (blue) on Wikidata, Mar 2017. Courtesy Vladimir Alexiev, Ontotext



Picture 21: Ghettos (orange) on Wikidata after EHRI intervention, Jul 2017. Courtesy Nancy Cooey, USHMM

Thesaurus Management System and Editorial Board

The EHRI-1 thesaurus was managed by ad-hoc means and by a small number of people. To make this activity more sustainable and facilitate the translation to a much larger number of languages, EHRI established an Editorial Board and deployed the VocBench[59] web-based thesaurus management system (TMS), which allows multiple users and a variety of roles.



antisemitismus, rasismus, fašismus a národní socialismus (cs); Antisemitismus, R
socialisme (fr); Antiszemitizmus, rasszizmus, fasizmus és nemzetiszocializmus (hu

armáda (cs); Militär (de); Military (en); militaire (fr); Katonaság (hu); Leger (nl); Strij

každodenní život (cs); Alltag (de); Daily life (en); vie quotidienne (fr); Mindennapi é

lidé (cs); Leute (de); Menschen (de); People (en); peuple (fr); Nép (hu); Mensen (nl);

nucené práce, deportace, gheta, tábory a konečné řešení táborů (cs); Zwangsarbei
et camp de la solution finale (fr); Kényszermunka, deportálás, gettók, táborok és megs
przymusowa, przesiedlenia, getta, obozy i obozy śmierci (pl); принудительный тр

období (cs); Periode (de); Zeit (de); Zeitabschnitt (de); Periods (en); période (fr); Idő

okupace, útlačování, diskriminace, pronásledování (cs); Besatzung, Unterdrückun
diszkrimináció, üldözés (hu); Megszállás, elnyomás, megkülönböztetés, üldözés (hu);

organizace (cs); Organisation (de); Organisations (en); organisation (fr); Szervezet

politika a ekonomika (cs); Politik und Wirtschaft (de); Politics and economics (en);

pomoc, bezpečí, záchrana (cs); Unterstützung, Wohlfahrt, Rettung (de); Aid, welfar
помощь, социальное обеспечение, спасение (ru)

reakce na okupaci a diskriminaci (cs); Reaktion auf Besatzung und Diskriminierun
discriminatie (nl); reakcja na okupacje i prześladowania (pl); ответ на оккупацию

státní aparát (cs); Staatsapparat (de); State Apparatus (en); appareil d'Etat (fr); dipo

Třetí říše (cs); das Dritte Reich (de); Drittes Reich (de); NS-Deutschland (de); NS-Staat
nazistowskie (pl); Trzecia Rzesza (pl); нацистская Германия (ru); Третий рейх (ru)

umění kultura věda náboženství (cs); Kunst, Kultur, Wissenschaft, Religion (de); A
наука религия (ru)

Židé a život Židů (cs); Juden, Jüdinnen und jüdisches Leben (de); Jews and Jewis

Fascism, and National Socialism (en); antisémitisme racisme , fascisme et national-
arodowy socjalizm (pl); антисемитизм, расизм, фашизм и национал-социализм (ru)

ettos, Camps and Final Solution Camps (en); travail forcé, déportation , ghetto , camp
wangarbeid, deportatie, getto's, kampen en vernietigingskampen (nl); praca

occupation , oppression , discrimination et persécution (fr); Megszállás, elnyomás,
owanie (pl); оккупация, гнёт, дискриминация, преследования (ru)

pl); политика и экономика (ru)

(nl); pomoc, opieka, ratownictwo (pl); помощь, благотворительность, спасение (ru);

(fr); Válaszok a megszállásra és diszkriminációra (hu); Reacties op bezetting en

(pl); государственный аппарат (ru)

dalom (hu); Náci Németország (hu); Derde Rijk (nl); Nazi-Duitsland (nl); Niemcy

); Kunst cultuur wetenschap religie (nl); sztuka, kultura, nauka, religia (pl); искусство

ven (nl); Żydzi i życie żydowskie (pl); евреи и еврейская жизнь (ru)

Select languages to display

Use this area to select languages you wish to see when viewing **concepts**, **categories** and **relationships** and to **use in searching** in your **current session.**

By default in every next login, the content language will be set to the users languages. All the changes in the content language will be valid for the current session only.

You will **always see only in user languages** when viewing **terms** associated with concepts and categories. To see in all the language, use the preferences area to check "Show only terms in selected languages" under "Administration" at top of the page.

Add new language

Order	Language (local)	Language (english)	Code
7	□ arabic	arabic	ar
46	□ বাংলা	bengali	bn
43	□ bosanski jezik	bosnian	bs
38	□ català	catalan	ca
13	☑ cesky	czech	cs
40	☑ dansk	danish	da
19	☑ deutsch	german	de
48	□ greek	greek	el
1	☑ english	english	en
4	□ español	spanish	es
15	□ persian	also known as farsi.	fa
32	□ suomi, suomen kieli	finnish	fi
41	□ vosa vakaviti	fijian	fj

○ Select all ○ Clear all Save Cancel

Legend 🟪 Proposed 🟩 Validated 🟦 Published 🟥 Revised ▨ Proposed deprecated ■ ~~Deprecated~~

Show more

Picture 22: VocBench showing EHRI thesaurus concepts with multilingual labels. Courtesy Boyan Simeonov, Ontotext

[59] http://vocbench.uniroma2.it/

VocBench is the best open source TMS that works directly over SKOS RDF data. SKOS is a widely used ontology for representing thesauri (see SKOS Primer,[60] SKOS Reference,[61] and 1.). VocBench works directly over Ontotext GraphDB and is used by UN FAO, EC's Publication Office, etc.

Access Points

The current version of the EHRI portal has full-text search. This can find documents by EAD access points (a.p.) mentioned in the document. However, it has the following deficiencies:

- The same concept may be spelt in many different ways (see below)
- It doesn't know multilingual labels of the concepts
- It doesn't know hierarchical relations: Eg. if a document mentions Lodz, it won't be returned in a search for Poland.

There is a staggering variety of writing out the same concept across institutions. Eg below are different ways of spelling Łódź (note: the first couple use Unicode combining diacritics instead of simple accented chars). Some of them are mapped to EHRI-1 access point objects, others are plain text:

Łódź (Ghetto)
Łódź (Poland)
Łódź (Poland),.
Łódz (Poland)
Łódź ehri-ghettos-513 cvocConcept
Łódź jc-places-place-iti-48 cvocConcept
Łódź terezin-places-place-iti-48 cvocConcept
Łódź (Poland)
Łódź – Łódź – Polen Łódź – Łódź – Poland
Lodz
Lodz terezin-places-place-iti-412 cvocConcept
Lodz (Polen)
Lodz - Poland
Lodz Ghetto
Lodz ghetto
Lodz ghetto, Poland
Lodz, Poland, Eastern Europe,
Lodz,Ghetto,Poland
Lodz,Lodz,Lodz,Poland
Lodz,גטו,Poland (note: "גטו" means Ghetto)

[60] https://www.w3.org/TR/skos-primer

[61] https://www.w3.org/TR/skos-reference

Lodž		
Lodž	jc-places-place-iti-48	cvocConcept
Lodž	terezin-places-place-iti-48	cvocConcept
Ladzogne		
Leitzmenstadt		
Lidzmanstadt		
Litmannstadt		
Litrwanstach		
Litsmannstadt		
Litz		
Litzen		
Litzm		
Litzmandstadt		
Litzmannsstadt		
Litzmannst		
Litzmannstadrt		
Litzmannstadt		
Litzmannstadten		
Litzmansntadt		
Litzmanst		
Litzmanstadt		
Litzmaustadt		
Lizmannstadt		
Lodcsh		
Lodez		
Lodezogne		
Lodon		
Lodsch		
Lodyogne		
Lodz		

Lodzen

Lodzi

Lodzogne

Losch

Lotdz

Lotz

Lotzogne

Lòdz

Lód

Lódz

Lódzen

Lödz

lodz

We have assembled a whole catalog of errors. A few of them are shown below:

- Lack of capitalization: maingain, olivier; maingain, roger; maires– belgique; maires--mechelen (belgique).
- Lot of mis-typed a.p., e.g. these should be **corporateAccess** instead of **subjectAccess**: 8f(United Nations Educational, Scientific & Cultural Organization [UNESCO]) 8g(Organisation f. Ernährung u. Landwirtschaft [FAO]). These should be **placeAccess** instead of **subjectAccess**: 915 (Balta, Chersson), 915 (Feodosia), 93. Koblenz.
- Compound a.p. that should be split up and applied as separate access point: Abortion – Poland--Oswiecim. Abortion– Poland--Warsaw. Abortion– Poland. Abortion.
- Wrong info (system of arrangement) mis-represented as **subjectAccess:** 5(dt. Orte in alphabet. Folge); 5(dt. Orte, alphabetisch); A-Z.
- Just a number without any text, i.e. an access point with no information content: 1; 100; 104. ID; 1042.

We used the Geonames co-referencing service (see 2.2.1 Geo-Referencing Service) to match geographic access points to Geonames, thus bringing multilingual access, unification of a variety of a.p. to the same entity, and hierarchical place access.

Other kinds of access points (subject, person, family, corporate) are still an open challenge. We have investigated the feasibility of co-referencing Person access points to VIAF and USHMM Persons, but research is still ongoing.

The challenges for corporate a.p. are especially hard, since it's hard to recognize them as designating an organization. E.g. consider this **subjectAccess** a.p. 8. Florian Geyer

This must be the organization 8th SS Cavalry Division Florian Geyer and not the person Florian Geyer (Franconian nobleman, diplomat, and knight), after whom the SS division was named. But because the a.p. is mis-typed (it should be **corporateAccess** not **subjectAccess**), that is hard to classify/interpret.

Conclusions

We presented some of the technical work in the EHRI project that is centered around the use of semantic and NLP technologies to help archivists and researchers working in the Holocaust domain. By semantic interlinking of information coming from a number of archives, structured domain databases (where available) and Linked Open Data, we can start building more complete histories of people involved in the Holocaust, their networks, and the events, places and times they were involved with.

Many challenges still remain since the amount of information (especially about people) is staggering. Collaboration with a wider community of LOD enthusiasts, genealogists and Holocaust researchers (including amateur researchers) will make it possible to correlate and interlink bigger amounts of information, making possible new kinds of research.

Acknowledgements

This work was performed in the EHRI-2 project, which is funded in part by the European Commission as H2020 project 654164. This paper is a greatly extended version of a presentation delivered at the workshop Data Sharing, Holocaust Documentation and the Digital Humanities. Best Practices, Benefits, Case Studies (DSDH 2017), June 2017, Venice, Italy.

We are thankful to all our colleagues in EHRI for their hard work and providing fruitful discussions and inspiration. The following organizations have participated in the technical work described in this paper: DANS, Ontotext, INRIA, KCL, NIOD, Cegesoma/Kazerne Dossin, Yad Vashem, USHMM, CDEC.

References

1. Baker, Thomas, Sean Bechhofer, Antoine Isaac, Alistair Miles, Guus Schreiber, Ed Summers. 2013. "Key Choices in the Design of Simple Knowledge Organization System (SKOS)." Journal of Web Semantics, 20 (May 2013): 35-49. DOI: 10.1016/j.websem.2013.05.001

2. Bennett, Giles. 2012. "The European Holocaust Research Infrastructure (EHRI)." *The Bulletin of the Carolyn and Leonard Miller Center for Holocaust Studies*, 16 (Spring 2012): 19. https://www.uvm.edu/sites/default/files/media/bull-2012.pdf

3. Blanke, Tobias, Michael Bryant, and Reto Speck. 2015. "Developing the collection graph." *Library Hi Tech*, 33/1 (2015): 610-623. DOI 10.1108/LHT-07-2015-0070

4. Brazzo, Laura, Silvia Mazzini. Open Memory Project. April 2015. https://www.bygle.net/wp-content/uploads/2015/04/Open-Memory-Project_3-1.pdf

5. Bryant, Michael, Linda Reijnhoudt, Reto Speck, Tibauld Clérice and Tobias Blanke." 2014. The EHRI Project - Virtual Collections Revisited." *HistoInformatics2014 - the 2nd International Workshop on Computational History, Barcelona 2014.* LNCS 8852 (2014): 294-303. DOI 10.1007/978-3-319-15168-7

6. Cooey, Nancy. 2018. "Using Wikidata to build an authority list of Holocaust-era ghettos." EHRI Document Blog. February 2018. https://blog.ehri-project.eu/2018/02/12/using-wikidata/

7. De Groot, Michael. 2010. "Building the New Order: 1938-1945." Spatial History Lab. Stanford University, 24 August 2010. https://web.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=51

8. EAD Converter User Guide, February 3, 2017, Ontotext Corp

9. Frankl, Michal. 2016. "Reports from the No Man's Land." *EHRI Document Blog*, 19 January 2016. https://blog.ehri-project.eu/2016/01/19/reports-from-the-no-mans-land/

10. Kahn, Rebecca. 2011. "The EHRI Project: building an online archive for European Holocaust Research." SCONUL Focus, 52 (2011):21-22. https://ehri-project.eu/sites/default/files/downloads/ehri_downloads/EHRI%20presentations/Sconul%20Focus%20Autumn%202011.pdf

11. Le Boeuf, Patrick, Martin Doerr, Christian Emil Ore, Stephen Stead. 2015. "Definition of the CIDOC Conceptual Reference Model, version 6.2.1." October 2015. http://www.cidoc-crm.org/.

12. Shoah Vocabulary Specification Beta Version, Fondazione Centro di Documentazione Ebraica Contemporanea, Regesta.exe. http://dati.cdec.it/lod/shoah/reference-document.html

13. Svensson, Lars. 2011. *AgRelOn, an Agent Relationship Ontology*. German National Library. Version 0.9, 30 June 2011. http://d-nb.info/standards/elementset/agrelon.html; http://d-nb.info/standards/elementset/agrelon.rdf; http://eudat.eu/sites/default/files/LarsSvensson.pdf.

14. Vanden Daelen, Veerle, Jennifer Edmond, Petra Links, Mike Priddy, Vaclav Tollar and Annelies Van Nispen. 2015. "Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives." HAL – Inria, December 2015. https://hal.inria.fr/hal-01281442v2

Last consultation URLs: 2019, February 3.