

# **Semantic Archive Integration for Holocaust Research: the EHRI Research Infrastructure**

**Vladimir Alexiev & Ivelina Nikolova**

**Ontotext Corp**

**DSDH 2017, Venice, 30 June 2017**



# Ontotext

- Started in 2000 as a Semantic Web pioneer
  - As Innovation lab within Sirma Group (listed as SKK), the biggest Bulgarian software house
  - Got spun-off and took VC investment in 2008
  - 65 staff, HQ in Bulgaria, reps in Canada, UK, Germany and USA
  - Over 400 person-years invested in R&D
- Multiple innovation & technology awards:
  - Washington Post, BBC, FT, BAIT, etc.
- Member of multiple industry bodies:
  - W3C, EDMC, ODI, LDBC, STI, DBpedia Foundation

# Ontotext Clients (selection)



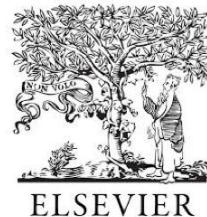
The Telegraph



OXFORD  
UNIVERSITY PRESS

THE  
BRITISH  
MUSEUM

HOUSES OF PARLIAMENT  
INFORMATION & COMMUNICATIONS TECHNOLOGY



WILEY  
Publishers Since 1807



IET  
The Institution of  
Engineering and Technology

tagasauris™



STANDARD  
& POOR'S

SPRINGER  
NATURE

babylon  
Everyone's personal health service.

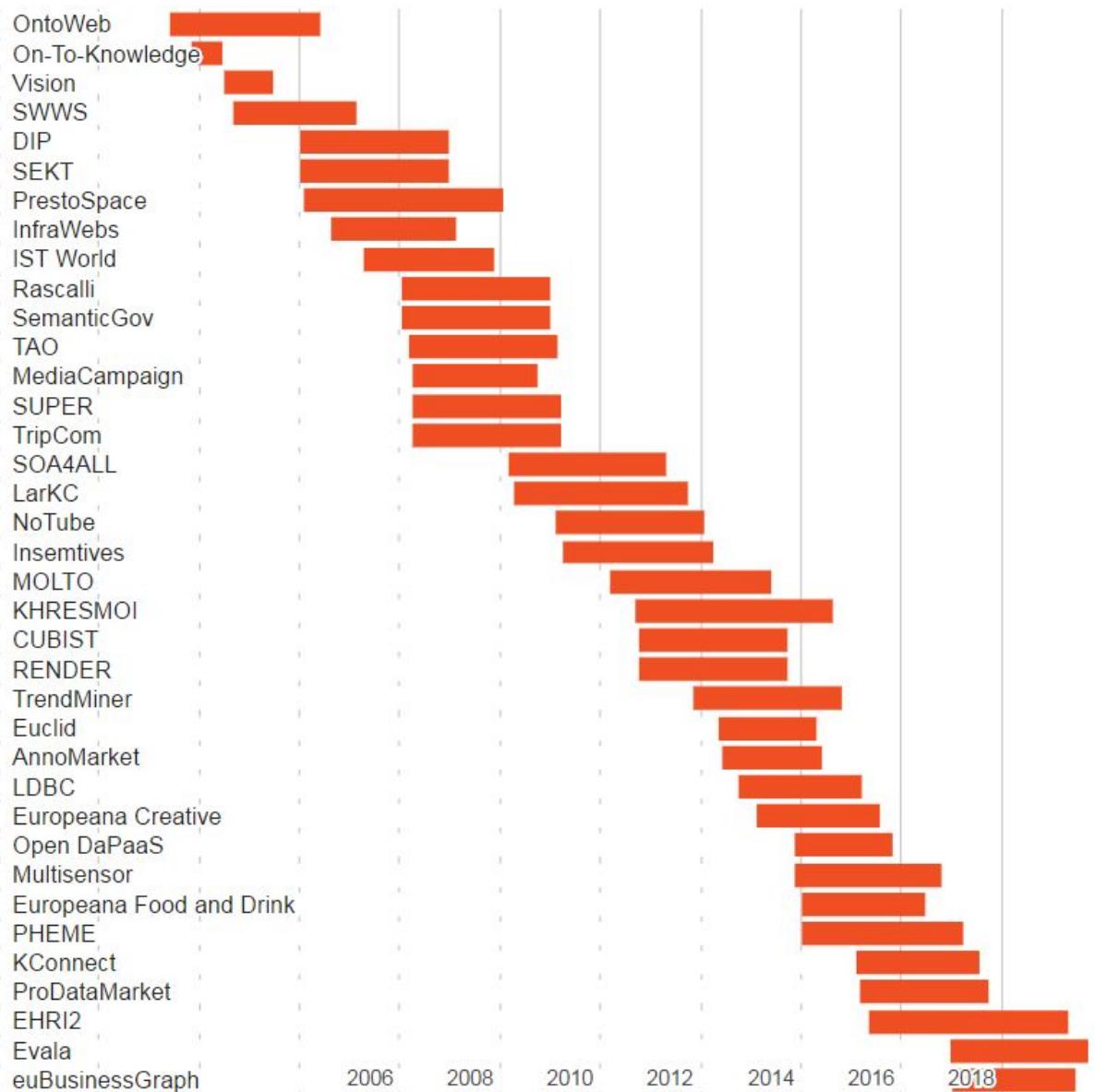
AstraZeneca  
Zeneca

FOUNDATION  
MEDICINE

syngenta



# Ontotext Innovation & Consulting



- 37 EC research projects
  - Bulgaria's most successful participant in the Framework Programmes
  - 5-7 active at any time
- Semantic technology
  - Semantic databases
  - Semantic enrichment
  - Multilingual translation
  - Media, news, sentiments, rumors, diversity
  - Life Sciences
  - Data and enrichment marketplaces
  - Commercial data and applications
  - eGovernment
  - LOD curriculum (education)
  - Cultural heritage

# EHRI Technical Work

- Main tech partners
  - KCL: portal development and operation, graph database
  - YV: thesauri and authorities
  - DANS: data architecture, EAD mapping, EAD transport
  - INRIA: standardization: EAD XSD, Schematron, TEI ODD, URLs
  - ONTO: EAD conversion, ingest, TMS, authorities, semantic enrichment, semantic search, LOD
- EHRI2 technical work packages
  - EAD (WP10) - converts archival descriptions from various formats to standard EAD XML
  - Authorities & Standards (WP11) - consolidates and enlarges EHRI authorities
  - Data infrastructures (WP13) - deduplication, semantic data integration, semantic text analysis.
  - Digital Historiography Research (WP14) - researcher tools based on semantic analysis, including Prosopographical approaches.

# EHRI Research Use Cases

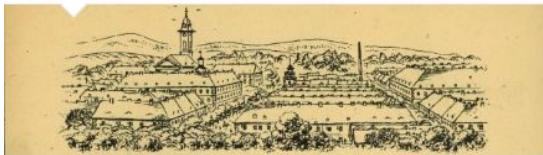
- Names and Networks: **Chances of Survival during the Holocaust.** Most Jews needed the support of other people to survive. Persons found aiders inside their personal and group networks. Investigate the networks in which European Jews operated during their persecution in the Second World War, and improve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

# EHRI Research Use Cases

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

# EHRI Document Blog

## Daily Orders from the Terezín (Theresienstadt) Ghetto

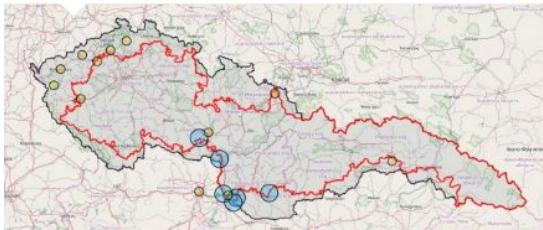


Daily Orders from the Terezín (Theresienstadt) Ghetto During the Second World War the Terezín/Theresienstadt Ghetto was one of the major sites of suffering and death for the Jews of the Bohemian Lands and several European countries including Germany, Austria, Netherlands, Denmark,...

[Continue Reading →](#)

⌚ 2016/05/18

## Reports from the No Man's Land



## Hans Frank – Letters from Exile

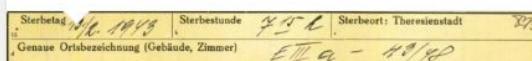


Correspondence of Hans Frank After the war, the Hansen family in Denmark found a suitcase with personal belongings of Hans Frank including correspondence and schoolbooks, brochures, annual school reports etc. In 1995, the Hansen family decided to donate the documents...

[Continue Reading →](#)

⌚ 2016/04/04

## Death Certificate of Gabriel Frankl from the Terezín Ghetto



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

## Testimony of Valerie Straussová



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

## 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 →](#)

- Provides inspiration for:

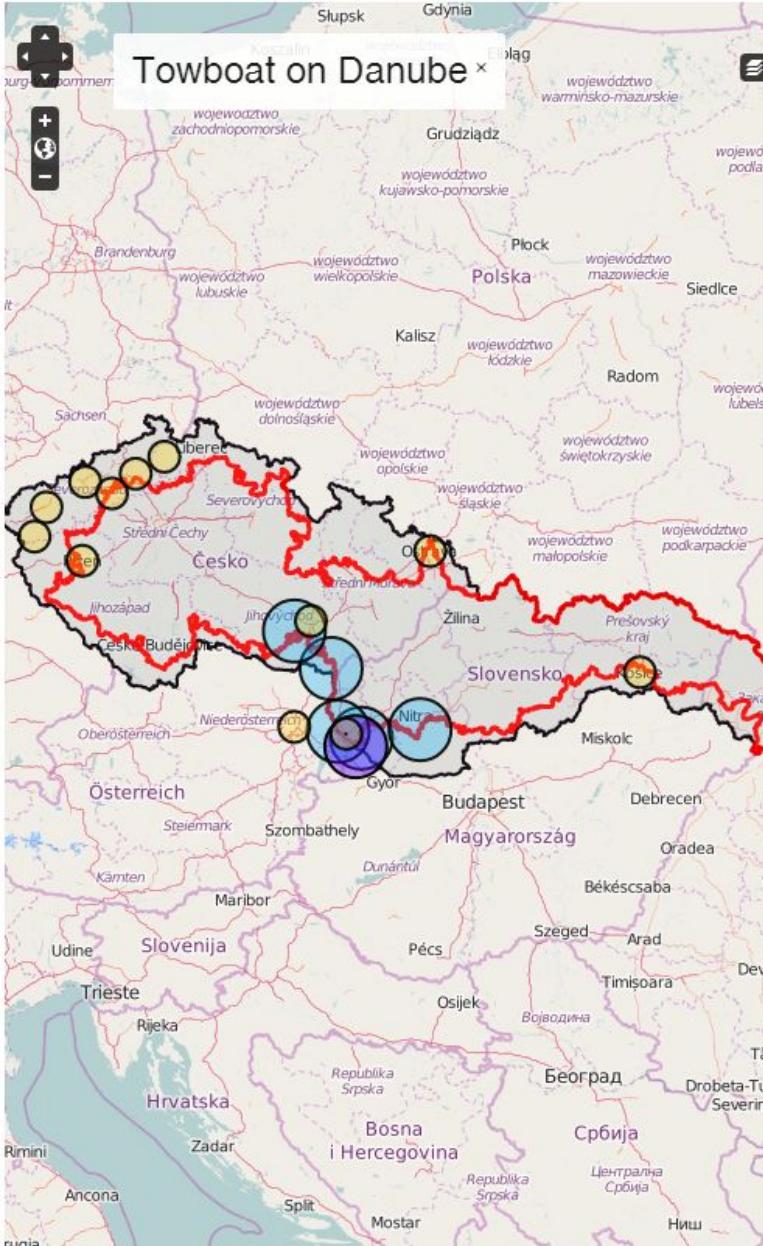
- Research questions
- Research techniques

- Our task:

- Enable such tools at scale
- Help researchers use them

- Soliciting other interesting examples!

# Geographic Mapping Example



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
2. Legal consequences in order

"Reports from the No Man's Land". EHRI Document Blog, 2016. Courtesy Michal Frankl, Jewish Museum in Prague and EHRI

# Controlled Access Points

- Some stats from end-2015

- 111k document units, 79k (70%) have access points
- number of access points: 686k (8.7x per object)
- 17% of a.p. instances have an object (link), 83% are just text
- Red represent errors, eg link=corporateBodyAccess should point to object=historicalAgent (or NONE), not to cvocConcept (concepts are not corporate bodies)

access point instances	object										
Subject / Link	country	cvocConcept	documentaryUnit	historicalAgent	repository	NONE	Total	Perc	Perc Obj		
cvocConcept						6	6	0.00%			
owl:sameAs						6	6	0.00%			
documentaryUnit	295	93034	39	22848	5	567424	683645	99.58%	17.00%		
corporateBodyAccess		9230	30	343	4	31786	41393	6.03%	23.21%		
creatorAccess		4		2984	1	23523	26512	3.86%	11.27%		
familyAccess		1	1	5		77	84	0.01%	8.33%		
genreAccess						6213	6213	0.90%	0.00%		
otherAccess	2	37	3	2		8	52	0.01%	84.62%		
personAccess				19497		62138	81635	11.89%	23.88%		
placeAccess	293	32300				126991	159584	23.24%	20.42%		
subjectAccess		51462	5	17		316688	368172	53.63%	13.98%		
historicalAgent		1813		504		564	2881	0.42%	80.42%		
associative				358			358	0.05%	100.00%		
hierarchical				74			74	0.01%	100.00%		
otherAccess				31			31	0.00%	100.00%		
subjectAccess		1813				564	2377	0.35%	76.27%		
temporal				41			41	0.01%	100.00%		
Total	295	94847	39	23352	5	567994	686532		17.27%		
Percent		0.04%	13.82%	0.01%	3.40%	0.00%	82.73%				

# Controlled Access Point Problems

- 23 ways of spelling Łódź across institutions.
  - Some of them are mapped to EHRI1 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,Łoż,Poland (note: “Łoż” means Ghetto)

Lodž

Lodž jc-places-place-iti-48 cvocConcept

Lodž terezin-places-place-iti-48 cvocConcept

# Controlled Access Point Problems

- E.g. leading spaces, weird punctuation, local IDs
  - YV uses <> as a marker for “missing place component”

Amtsgericht Welun  
\*Frankfurt/Main  
.SUCHY BARBARA, DR  
o(Außenpolitik)  
<> Friesland <> The Netherlands placeAccess  
Sofia Sofia <> Bulgaria placeAccess  
<> <> <> Bulgaria placeAccess  
(Kleinmann Ferdinand  
(Pervomaisk (Olviopol, Bogopol, Golta  
""She'erith Hapletah""  
-9 Hitler als Staatsmann und Politiker  
1 (Kirchen, Pazifisten, Sozialisten, Sonstige allg.)  
1(Fürstenenteignung

- Empty a.p. or only an ID

-  
.1  
100.  
104. ID

# Controlled Access Point Problems

- Mis-represented (mis-typed) access points:

1941	subjectAccess	# it's just year
1942	placeAccess	# it's just year
5(dt. Orte in alphabet. Folge)	subjectAccess	# System of Arrangement
A-Z	subjectAccess	# System of Arrangement
8f(United Nations Educational, Scientific & Cultural Organization [UNESCO])	subjectAccess	# corporateAccess
8g(Organisation f. Ernährung u. Landwirtschaft [FAO])	subjectAccess	# corporateAccess
915 (Balta, Chersson)	subjectAccess	# placeAccess
915 (Feodosia)	subjectAccess	# placeAccess
93. Koblenz	subjectAccess	# placeAccess
8. Florian Geyer	subjectAccess	<u># corporateAccess: 8th SS Cavalry Division Florian Geyer</u>

- Compound (pre-coordinated) access points. Split up and match separately

Abortion--Poland--Oswiecim.	subjectAccess	# but there's also placeAccess
Abortion--Poland--Warsaw.	subjectAccess	# but there's also placeAccess
Abortion--Poland.	subjectAccess	# but there's also placeAccess
Abortion.	subjectAccess	# no relation between this and above 3

- Compound a.p. that should have been entered as separate a.p. entries

Aachen, Arnsberg, Baden, Bayern, Berlin, Bremen, Düsseldorf subjectAccess

# Building Up Authorities

- "Controlled" Access Points? Not at all, this is a misnomer
- Negative impact on:
  - Search (user can't know what to search for)
  - Indexing (when creating a link on the portal, indexers not sure what to select)
- Need to harmonize at the portal and build up authorities.
  - Unicode normalization
  - Removing punctuation and other cleanup
  - Splitting up compounds (atomization)
  - Heuristics to treat a.p. type: can't trust it completely, but shouldn't ignore it
- Status per kind
  - Places: strong advancement, used Geonames as reference database (80k out of 10M)
  - Concepts: EHRI Thesaurus is made sustainable through a TMS and Editorial Board, but IMHO small (2-5k)
  - Ghettos & Camps: matching to Wikidata to bring in more data and harmonize
  - Persons: evaluating VIAF and USHMM databases for coverage
  - Organizations: should start by harvesting corporateAccess across providers

# Authorities-Related Tasks

- Consolidation and enlargement of EHRI authorities
  - To make indexing and retrieval of information more effective.
- Access Points in ingested EADs
  - Normalization of Unicode, spelling differences, punctuation; deduplication; clustering; coreferencing to global authorities where available
- Enlarging and integrating disparate EHRI thesauri
  - Subjects: deployment of a Thesaurus Management System, editorial board, deciding whether "candidate" concepts harvested from access points should be added to authorities
  - Places: coreferencing to Geonames
  - Camps and Ghettos: integrating data from Wikidata
  - Persons, Corporate Bodies: starting
- Implementation of semantic (conceptual) search
  - Including hierarchical query expansion
  - Aiding indexing for better interconnectivity of archival descriptions
- Use a mix of automatic and manual approaches

# Geo-Referencing Service

- Uses Geonames to find place references
  - In text (oral history), access points, local databases (USHMM Places)
- Based on Ontotext commercial application pipelines
- Specially tailored matching guidelines. E.g. not a place:
  - 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: publication
  - USS [America]: ship
- Remove place types to increase recall and avoid false matches:
  - Kreis \* (District in German)
  - \*falu (village in Hungarian)
  - Ghetto (numerous villages in Italy)
  - Canton (Swiss/French region, matches old name of Guangzhou, China)
- Prefer bigger cities, use the place hierarchy for context
- Hybrid application based on Machine Learning and refinement rules

# Geo-Referencing Service

- Sophisticated disambiguation mechanisms are developed, based on the following place characteristics
  - places names in various languages - Oświęcim, Освенцим, Auschwitz,
  - synonym labels - Auschwitz-Birkenau, Birkenau
  - place hierarchy
    - *Moscow (Russia)*: as opposed to the 23 places of the same name in the US, and a few more in other countries
    - *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 Gold Standard over news corpora
  - for populated places, we give priority to bigger places

# Geo-Referencing Service

- 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.
- Evaluation on 500 access points

Type	Precision	Recall	F-measure
placeAccess	0.97	0.86	0.91
subjectAccess	0.88	0.90	0.89

# Geo-Referencing Service



*Map of places extracted from a  
USHMM Oral History interview.  
Courtesy Tobias Blanke, King's  
College London, 2016*

# Geographic Challenges

- Geonames often includes historic place names
  - Even historic countries (e.g. Czechoslovakia)
- Some challenges of Geonames for historical geography
  - Nazis renamed place names,  
e.g. *Oświęcim* → *Auschwitz*, *Brzezinka* → *Birkenau*.
  - Nazis established new administrative districts,  
e.g. Reichskommissariat Ostland included *Estonia*, *Latvia*, *Lithuania*, the northeastern part of *Poland* and the west part of the *Belarusian USSR*
  - Historic processes changed borders, place names and administrative subordination.  
e.g. Wilno (Vilna) was part of Poland until 1939, when USSR gave it to Lithuania, to become its capital Vilnius.
  - Geonames place hierarchy reflects modern geography  
e.g. "Wilna--Poland" disambiguated as two places "Vilnius, Lithuania" and "Poland"
- Fixes?
  - Local Geonames fixes (e.g. North America < Americas the megacontinent, not the Cuban village)
  - Considering local Geonames additions (e.g. Czechoslovakia the parent of Czech Republic and Slovakia)
  - Considering other sources of historical geography (e.g. Spatial History Project)

# Subjects/Concepts: VocBench TMS

- VocBench: Best open source TMS

- Runs directly over semantic data (ONTO GraphDB); partnership with developer (UniRoma2)
- Used by EC OPOCE (EuroVoc), UN FAO (AgroVoc)
- E.g. multilingual labels

The screenshot shows the VocBench web application interface. At the top, there's a navigation bar with links for 'Global data management', 'Administration', 'About VocBench', 'English', 'RSS feed', 'Preferences', 'Help', and 'Sign out'. Below the navigation bar, there's a search bar with dropdowns for 'Exact word' and 'Go'.

The main content area has tabs for 'Concepts', 'Properties', 'Schemes', 'SPARQL', 'ICV', 'Validation', and 'History'. The 'Concepts' tab is selected. On the left, there's a sidebar titled 'Concepts' with a list of terms like 'antisemitismus', 'armada', 'každodenní život', 'lidé', 'nučené práce', 'období', 'okupace', 'organizace', 'politika a ekonomika', 'pomoc', 'reakce na okupaci', 'státní aparát', 'Třetí říše', 'umění kultura věda náboženství', and 'Židé a život Židů'. To the right of the sidebar is a 'Select languages to display' dialog box. This dialog box contains instructions about selecting languages for the current session, a note about defaulting to user languages, and a table for adding new languages. The table lists various languages with their local names and English names, along with checkboxes and edit/cancel buttons. Languages listed include arabic, bengali, bosnian, catalan, czech, dansk, deutsch, greek, english, español, persian, suomi, suomen kieli, and filian. At the bottom of the dialog box are 'Select all' and 'Clear all' buttons, and 'Save' and 'Cancel' buttons.

The main right-hand panel displays search results for terms like 'Fascism, and National Socialism', 'ghettos, Camps and Final Solution Camps', 'occupation , oppression , discrimination et persécution', 'reakce na okupaci a diskriminaci', 'státní aparát (cs)', 'Třetí říše (cs)', 'umění kultura věda náboženství (cs)', and 'Židé a život Židů (cs)'. Each result is followed by its definition in multiple languages (e.g., English, French, German, Spanish, Russian).

At the bottom of the page, there's a legend: Proposed (pink square), Validated (green square), Published (blue square), Revised (red square), Proposed deprecated (gray square), and Deprecated (black square). There's also a 'Show more' button with a right-pointing arrow.

# Subjects/Concepts: VocBench

## - Concept information

Signed in as Administrator ( Administrator  ) : EHRI

Global data management | Administration | About VocBench | English | RSS feed | Preferences | Help | Sign out

VocBench VERSION 2.4.2 [Build 20160531] (DEVELOPMENT)

Exact word  Go Advanced search

Concepts Properties Schemes SPARQL ICV Validation History Concept navigation history Content language

Concepts Show URI Show non-preferred

URI: <http://data.ehri-project.eu/thesaurus/concept/379>

+ antisemitismus, rasismus, fašismus a národní socialismus (cs); Antisemitismus, Rassismus, Faschismus und Nationalsozialismus (de); Antisemitism (TEST) (en); Antisemitism, Racism, Fascism, and National Socialism (en); antisémítisme racisme , fascisme et national-socialisme (fr); Antiszemitzmus, rassizmus, fasizmus és nemzetiszocializmus (hu); Antisemitisme, racisme, fascisme en nationaal-socialisme (nl); antysemifyzm, rasizm, faszyzm oraz narodowy socjalizm (pl); антисемитизм, расизм, фашизм и национал-социализм (ru)

+ armáda (cs); Militär (de); Military (en); militaire (fr); Katonaság (hu); Leger (nl); Strijdkrachten (nl); Sily zbrojne (pl); wojsko (pl); военные (ru)

+ každodenní život (cs); Alltag (de); Daily life (en); vie quotidienne (fr); Mindennapi élet (hu); Dagelijks leven (nl); życie codzienne (pl); повседневная жизнь (ru)

+ lidé (cs); Leute (de); Menschen (de); People (en); peuple (fr); Nép (hu); Mensen (nl); ludzie (pl); люди (ru)

+ nucené práce, deportace, gheta, tábory a konečné řešení táborů (cs); Zwangarbeit, Deportation, Gettos, Konzentrationslager und Vernichtungslager (de); Forced labour, Deportation, Ghettos, Camps and Final Solution Camps (en); travail forcé, déportation , ghetto , camp et camp de la solution finale (fr); Kényszemunka, deportálás, gettók, táborok és megsemmisítő táborok (hu); Kényszermunka, deportálás, gettók, táborok és végső megoldás táborok (hu); Dwangarbeid, deportatie, getto's, kampen en vernietigingskampen (nl); praca przymusowa, przesiedlenia, getta, obozy i obozy śmierci (pl); принудительный труд, депортации, гетто, лагеря и лагеря окончательного решения (ru)

+ období (cs); Periode (de); Zeitabschnitt (de); Periods (en); période (fr); Időszakok (hu); Tijdsperioden (nl); okresy (historii) (pl); периоды (ru)

+ okupace, útlačování, diskriminace, pronásledování (cs); Besatzung, Unterdrückung, Diskriminierung und Verfolgung (de); Occupation, Oppression, Discrimination and Persecution (en); occupation , oppression , discrimination et persécution (fr); Megszállás, elnyomás, diszkrimináció, üldözés (hu); Megszállás, elnyomás, megkülönböztetés, üldözés (hu); Bezetting, onderdrukking, discriminatie en vervolging (nl); okupacja, opresja, dyskryminacja i prześladowanie (pl); оккупация, гнёт, дискриминация, преследование (ru)

+ organizace (cs); Organisation (de); Organisations (en); organisation (fr); Szervezetek (hu);

Legend Proposed Validated Published Revised Proposed deprecated Deprecated

antisemitismus, rasismus, fašismus a národní socialismus (cs); Antisemitismus, Rassismus, Faschismus und Nationalsozialismus (de); Antisemitism (TEST) (en); Antisemitism, Racism, Fascism, and National Socialism (en); antisémítisme racisme , fascisme et national-socialisme (fr); Antiszemitzmus, rassizmus, fasizmus és nemzetiszocializmus (hu); Antisemitisme, racisme, fascisme en nationaal-socialisme (nl); antysemifyzm, rasizm, faszyzm oraz narodowy socjalizm (pl); антисемитизм, расизм, фашизм и национал-социализм (ru)

Show inferred and explicit Show/hide tabs

Terms (13) Definition (0) Attribute (0) Relationship (0) Alignment (0) Note (0) Image (0) Scheme (1) Hierarchy History

Language	Term
English (en)	<input checked="" type="checkbox"/> <input type="checkbox"/> Antisemitism, Racism, Fascism, and National Socialism (Preferred) W <input type="checkbox"/> Antisemitism (TEST) W
Français (fr)	<input checked="" type="checkbox"/> <input type="checkbox"/> antisémítisme racisme , fascisme et national-socialisme (Preferred) W
Русский (ru)	<input checked="" type="checkbox"/> <input type="checkbox"/> антисемитизм, расизм, фашизм и национал-социализм (Preferred) W
Cesky (cs)	<input checked="" type="checkbox"/> <input type="checkbox"/> antisemitismus, rasismus, fašismus a národní socialismus (Preferred) W
Deutsch (de)	<input checked="" type="checkbox"/> <input type="checkbox"/> Antisemitismus, Rassismus, Faschismus und Nationalsozialismus (Preferred) W
Hungarian (hu)	<input checked="" type="checkbox"/> <input type="checkbox"/> Antiszemitzmus, rassizmus, fasizmus és nemzetiszocializmus (Preferred) W
Polski (pl)	<input checked="" type="checkbox"/> <input type="checkbox"/> antysemifyzm, rasizm, faszyzm oraz narodowy socjalizm (Preferred) W
Nederlands (nl)	<input checked="" type="checkbox"/> <input type="checkbox"/> Antisemitisme, racisme, fascisme en nationaal-socialisme (Preferred) W
- (ru-latn)	<input type="checkbox"/> <input checked="" type="checkbox"/> antisemitism, racism, fascism and national-socialism (Preferred) W
- (uk)	<input type="checkbox"/> <input checked="" type="checkbox"/> Антисемитизм, расизм, фашизм та национал-соціалізм (Preferred) W
- (sh-latn)	<input type="checkbox"/> <input checked="" type="checkbox"/> antisemitizam, rasizam, fašizam i nacional-socijalizam (Preferred) W

Show more

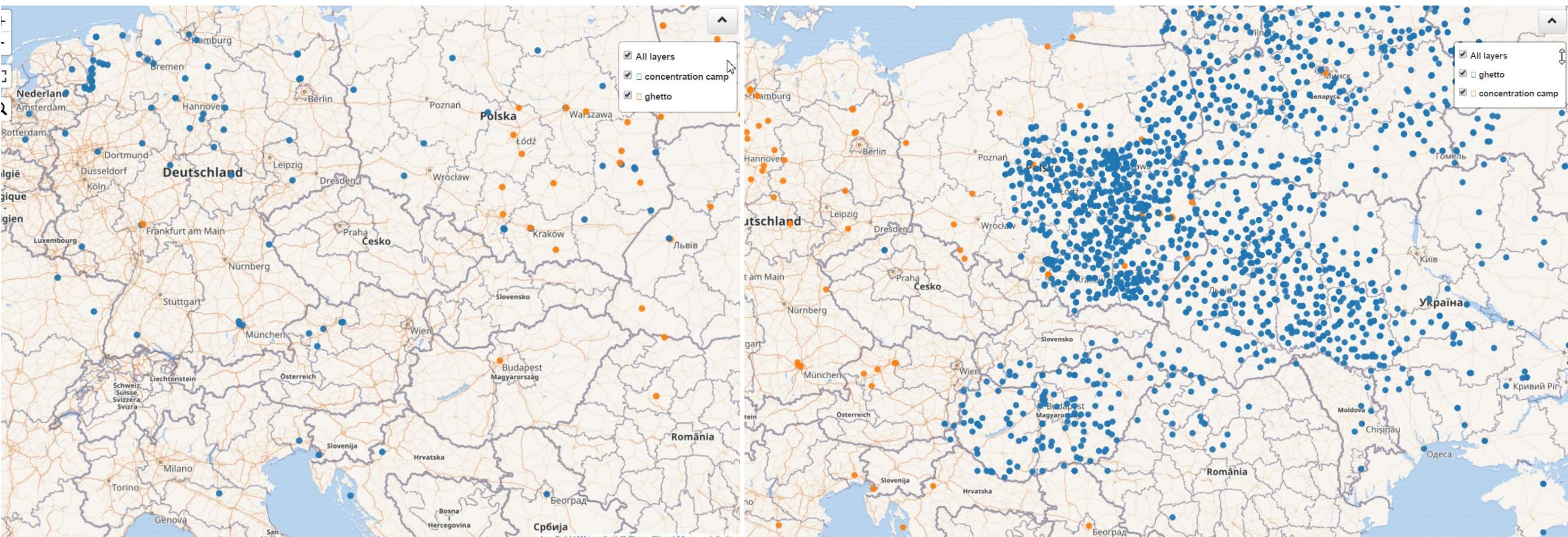
© FAO & ART Group , 2017

# Ghettos & Camps

- EHRI1 published Ghetto & Camp data, but was poor
  - E.g. camp [2030](#): only a label "Maly Trostinec"
- Wikidata knows:
  - 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, eg Vernichtungslager Maly Trostinez, KZ Maly Trostinez, Blagowschtschina
  - country: Belarus
  - location: 53°51'3"N, 27°42'17"E
  - Authority IDs: Geonames, VIAF, Freebase
- Wikipedia knows a lot more, but the info is not structured
  - names: Maly Trostinet, Maly Trastsianets, Trasciane, Малы Трасцянец, Maly Tras'tsyanyets, Малый Тростенец, Maly Trostinez, Maly Trostenez, Maly Trostinec, Klein Trostenez
  - Location, Nazi admin district, date established, victim countries, victim places of origin
  - killing grounds: Blagovshchina (Благовщина) forest, Shashkovka (Шашковка) forest
  - known victims, perpetrators (and their fate)

# Ghettos & Camps: Wikidata

- EHRI2 decided to coreference to Wikidata
  - So we can get lots more labels, coordinates, place/location...
  - And international collaboration! E.g. [Wikidata.GLAM Facebook Group](#)
  - [Camps and Ghettos on Wikidata](#): before (15 Mar 2017) and now (30 Jun 2017):



# Semantic Annotation (Text Enrichment)

GATE Developer 8.1 build 5169

File Options Tools Help

Messages RG-50.120.0268 ... AGNES MANDL ADA... Application

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

Microsoft Word - RG-50.120.0268\_tcn\_en.doc

RG-50.120 #0268 SOLI GANOR 1.01 Soli Ganor [nee Zelig, Hikind?] was born in a small town of Heydekrug in the border between Germany and Lithuania. Population spoke only German, as did Soli. His parents hailed from Kovno. 2.03 In 1933 Soli's father decided to sell his soap factory and moved back to Kovno. 2.14 Soli was born in 1928, on May 18. 2.29 In Kovno they had a big family, on the maternal side. Soli enjoyed being with their cousins and uncles. Jewish culture prospered. Yiddish prevailed there. 5.05 Soli's school days and happy childhood until 1939. 6.54 Soli's father had several small businesses and the family was well off. 7.49 Soli had a sister who was 14 years older, Fanya, and his brother was 7 years older, Zvi [Herman]. 8.29 Soli's paternal grandparents, from Minsk, were very observant, but Soli's father entered the Menshevik movement at age 13. Soli's family, therefore, was not religious. After 1917, Soli's father had a minor appointment in Krenski's government. [He had already met Trotsky and Lenin.] When the Bolsheviks won, Soli's father was sentenced to death, but one of his friends from youth, who was a Bolshevik, helped him flee to Lithuania. 11.21 Soli's father was not a Zionist; he was a convinced Socialist, and didn't see the future of the Jews in Zionism. Soli's mother was a Zionist, as were 95% of the Lithuanian Jews. 16.00 Soli's one year sojourn in a Lithuanian school. Transfer to Yauneh, an orthodox gymnasium, where he studied 4 years. 23.45 First echoes of war in Kovno: analysis of the situation and decision of the family about best strategy. 30.55 Family received a US visa in 1938, but they were afraid to start again, and sad to leave their large families. 34.51 In June 1940, the Russians entered Lithuania, and all hopes for moving out were squashed. 38.21 Meeting with the Japanese consul Sugihara at his aunt's gourmet store. Soli and Sugihara spoke Russian. [Great story!] Relationship with Sugihara. 47.27 June 1940, invasion of Russia into Lithuania. <http://collections.ushmm.org> Contact reference@ushmm.org for further information about this collection

2.21.49 Soli's family attempt to flee to Russia but without success. Father continues doing well in his business until 1941. Then the Russians began deporting people to Siberia. 26.49 Soli's older brother, working for the Russian intelligence, managed to get his family off the deportation lists several times. Brother was found out. The escape to Russia doesn't succeed. The return to Kovno. 3.29.53 They are caught a short distance from Kovno, and sent to the 7th Fort. 36.06 S's brother disappeared there. For some reason, S and his mother were returned home, and delivered to the door in Kovno. To their amazement, they found the father and the sister there. It was mid-July 1941. 45.44 The story of the toy wooden horse S got on his 4th birthday. 57.00 End of July, order is given to move into a ghetto in Vilijampole [Slobodka]. TAPE II 4.00.00 Friendship with Petras, the Lithuanian, makes it possible for S's family to exchange homes to enter the ghetto. 10.00 The move into the ghetto before it gets sealed; settling in. Work. 15.00.00 S and his father witness the burning of the infectious disease hospital and the small ghetto, with everyone inside, including Soli's aunt who was a nurse there. He and his father were also ordered to evacuate the orphans and other sick people onto trucks. This was the first Action. Those who were not burned alive were transported to be killed in the 9th Fort. 30.00 The second Selection. S is saved at the last minute. The evacuees are sent to the 9th Fort. 47.47 Eye witness to the murder in the 9th Fort. The teller: 'Cookie'. 6.13.26 The name was 'Cookie' Kopelman. 17.07 S works in the carpentry shop which prevents him from going to toll in the 'airport'. 19.07 Activities during that period, illegally reading, studying, with professor Edelstein[?]. 21.00 Valuables and books Actions. 33.44 Hiding the books. 43.50 Professor Edelstein is killed for holding a book. 45.30 Professor Chaim Perlman[?]. His influence on the youth. He died of hunger. <http://collections.ushmm.org> Contact reference@ushmm.org for further information about this collection

Annotations List

Annotations Stack

Co-reference Editor

Text

Event

Location

Organization

Person

RelationPersonRole

Species

Work

Original markups

PLO

PLO\_ANNIE

trusted

DC:CREATOR jvanopstal

DC:TITLE Microsoft Word

META:AUTHOR jvanopstal

MimeType application/pdf

TIKA\_DC:CREATOR jvanopstal

TIKA\_DC:TITLE Microsoft Word

TIKA\_META:AUTHOR jvanopstal

gate.SourceURL file:/home/ive...

sentimentScore 0.4983852157

Document Editor Initialisation Parameters Relation Viewer

- Entity extraction from free text
- GATE toolkit (University of Sheffield, ONTO is second largest contributor)
- Large KB gazetteers
- Disambiguation is very important
- Relies on the existence of KB with rich features (e.g. relations between entities)
- Enables semantic search and powerful research techniques
- Relation extraction can extract events & relations between people
- Will try for Jewish Council text

# Jewish Social Networks

- Tasks
  - Deduplication/coreferencing of person records (partial data!)
  - Reconstruct family and acquaintance relations
  - Do social network analysis
  - Answer hypotheses, e.g. on chances of survival
- Person Data Sources
  - USHMM Persons (HSV) database (3.2M+1M names, 1.2M places, 2.5M dates...):  
Coverage of Persons in Interviews is 15%
  - Dutch Jewish Digital Monument website (100k's names): may not get it
  - CDEC person database (10k's names): upcoming
  - Dutch Jewish Council proceedings: textual, not structured records; upcoming
  - Virtual International Authority File (VIAF) (15M): only "famous" people, potential poor coverage of Holocaust
  - Wikidata (2.5M): potential poor coverage of Holocaust domain

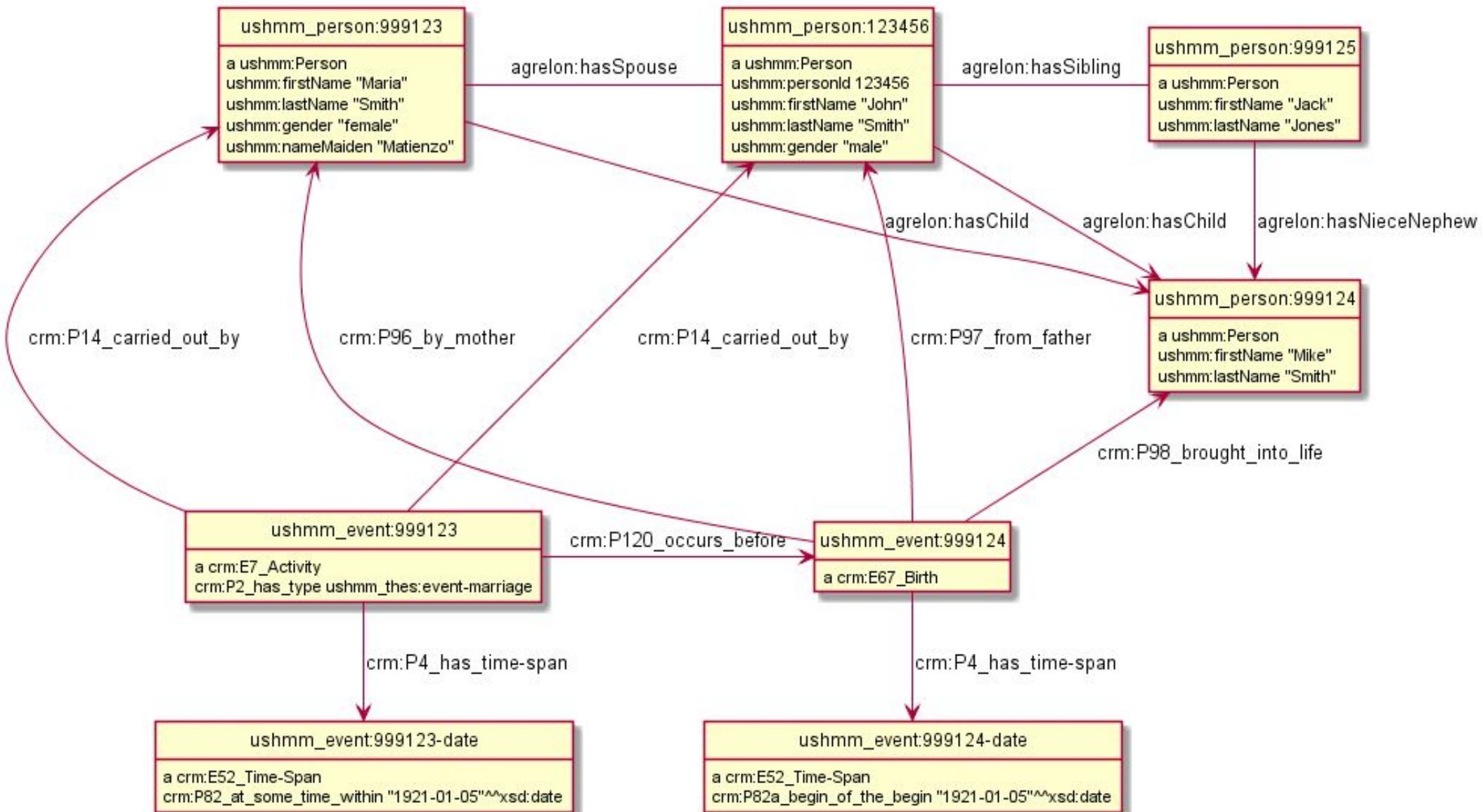
# Jewish Social Networks

- USHMM Survivors and Victims (persons) includes:
  - 3.2M person records (public part), 3M in a private/secured part
  - 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).
- Data came from 5-15k lists.
  - Some persons have 5-6-7 records

# Reconstructing Families

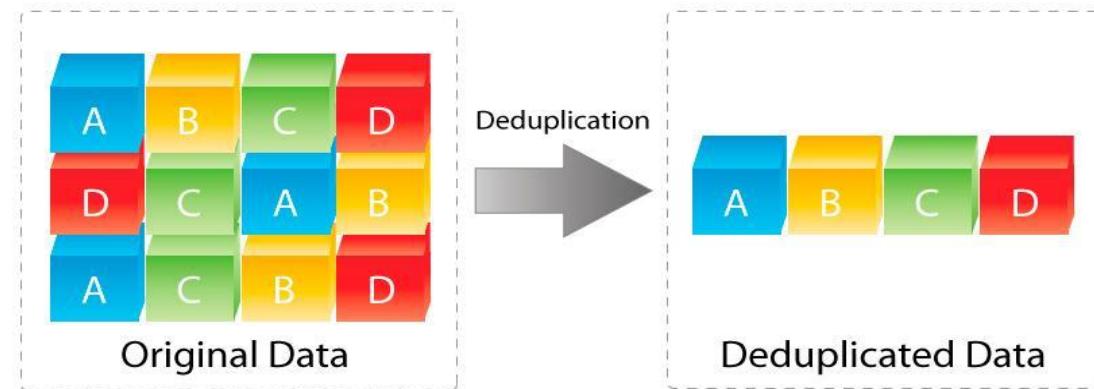
- Example person record with Additional names:
  - personId 123456: firstName “John”, lastName “Smith”, gender Male, nameSpouseMaiden “Matienzo”, nameSpouse “Maria Smith”, dateMarriage 1921-01-05, nameChild “Mike Smith”, nameSibling “Jack Jones”.
- We can convert "additional names" into a network of "stub" persons and events, and try to match to other person records
  - 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
- Next figure uses CIDOC CRM to represent persons and events, and AgRelOn to represent person relations

# Reconstructing Families



# Person/Family Identification: Judgements

- Gold Standard with human judgement about possible matches
  - Using name search and historic reasoning
  - Difficult because of incomplete info
- Then we train a Machine Learning model for matching



- 5.1k lists (sources of USHMM persons)
  - Some correspond to coherent dated historic events.
  - Can provide probable info about each person in the list
  - Copy "default values" (eg "child", "Polish") and dates & places to person
  - Provide evidence the persons listed in that source may have known each other

# Example of Historic Reasoning

- Sof'ia Alkhazova ([Latin](#), [Cyrillic](#)) was married to [MOISEI CHERNIAKOVSKII](#)
- But this can only be deduced from source [doc](#) that has extra info missing in structured list:
  - Evacuated from Odessa to Urgench (region Harezemsk, Uzbekistan): same place as husband
  - Had a son named Boris Moiseev Chernyakovskiy: same patronymic & family name

162

Фамилия <i>Алхазова</i>	Имя <i>София</i>
Отчество <i>Родионовна</i>	Отношение к главе семьи <i>жена</i>
Пол <i>ж</i>	Год рождения <i>1994</i>
Место рождения	
Специальность и стаж <i>зубн. вр</i>	Национальность <i>евр</i>
Местожительство до эвакуации: область (край) район	Где работал до эвакуации <i>Демихов</i>
город, село <i>Одесса</i>	Цех
Кем работал (должность) <i>брзр</i>	№ списка <i>114</i> стр. № по списку <i>904</i>
Где поселен (адрес) <i>ул. пр. Ургенч Комиссаров, 11</i>	Где работает в настоящее время Учрежд. Должн.
Правка выдана: кому огда из ЦСУ, № 38 7204 48 1-42 г. т. 25300.	
1942 г.	
Тип. НКВД, зак. № 514 25300 42 г.	

Список детей до 16-лет, проживающих вместе

Фамилия	Имя	Отчество	Отнош. к главе семьи	Возраст
<i>Черниковский</i>	<i>Борис</i>	<i>Моисеев</i>	<i>сын</i>	<i>1937</i>
2				
3				
4				
5				
6				

# Historical and Place Reasoning

- After coreferencing to Geonames, we can use place hierarchy and proximity (based on coordinates) to make likely inferences.
- E.g. OH interview [RG-50.661.0001](#):
  - "My Grandfather owned a bank in partnership with a cousin, Leibela Mandell"
  - "The family remained in Poland until the second world war"
  - Mentions Premishlan (Przemyslany in Polish)
- Search for [Lejbela](#) (First name, Soundex) and [Mandel](#) (Last name, exact) finds [Lejb Ela MANDEL](#).
- List [Initial Registration of Lublin's Jews - October 1939 and January 1940](#)
  - "listing of the male heads of households appears to have survived in the Lublin Judenrat files".
  - Represents 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 male: "second oldest son was Leib"
- Przemyslany is now Peremyshlyany, Lviv Oblast, Ukraine (not Poland).
  - But [Google Maps](#) shows the distance from Lublin to Peremyshlyany is 263 km
- So it is very likely that [Lejb Ela MANDEL](#) is the cousin mentioned in the interview.

# Place Hierarchy for Query Expansion

- Query Expansion means finding items (persons, docs)
  - Indexed by any sub-term (e.g. Amsterdam)
  - When the user searches for a super-term (e.g. Netherlands)
- We can 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 we can extract a lot more Netherlands-related persons, perhaps 100k

# Example of Semantic Faceted Search

- Hierarchical Facets: Food & Drink Topic, Geography

Fork me on GitHub



## Europeana Food & Drink

The Semantic Demonstrator shows the use of semantic technologies for classification and discovery of Europeana objects related to Food and Drink. Detailed description, data, SPARQL endpoint.

No active filters

ontotext made with europeana

View on map

Food and Drink

- + Agriculture 25266
- + Beverages 14856
- + Cuisine 26796
- + Eating behaviors 27721
- + Food and drink by country 10858
- + Food and drink preparation 28386
- + Food and drink terminology 10581
- + Food and the environment 1027
- + Food culture 2753
- + Food decorations 2166
- + Food festivals 6

Places

- + 10861432 1568
- + Africa 3776
- + Americas 92
- + Antarctica 4
- + Arabian Peninsula 33
- + Asia 2074
- + Atlantic Ocean 45
- + Australia (continent) 3
- + Bering Sea 1
- + British Empire 12
- + Community of Latin American and Caribbean States 203

Results per page: 24 ▾

Results 1 - 24 of 43685

◀ Page 1 of 1821 ▶



Artists barter pictures for food at Paris exhibition . The ' Barter Selon ', the art exhibition which comes to the rescue of



Fertility - c.1902 by Munch, Edvard (1863-1944) Location Nasjonalgalleriet, Oslo, Norway In The National Gallery of Oslo Edvard



Helen Wills drinks a toast . Mrs Helen Wills Moody , the American tennis ' Queen ' who is a regular spectator at the Davis Cup matches



Prancūžiškai pagaminti žirneliai (receptas)



0,5 litro šviežių arba šaldytų žirnelių, 1 šaukštės sviesto, 1 gūžė žalių salotų, druskos ir smulkintu krau. Žalumvns



Džiovintų vaisių dėserto gamyba (receptas)



Grilled aubergine rolls with herbs on wooden skewer credit: Ilva



Freshly rinsed apricots drying on checked tea towel credit: Lara



Traditional style fruit tea loaf sliced and served with butter and

# Person Deduplication

- Data preprocessing
  - Person Name Normalization, Daitch-Mokotoff encoding, Beider-Morse Phonetic (eg Maria Izsak -> MR ASSK)
  - Statistical and Rule Based models for assigning of person gender
    - Eg Maria is "female" 0.999 prob according to Linear Class algorithm and 0.9 according to Rule Based
  - Linking USHMM Places to Geonames: proximity indicates likelihood
- Statistical model for automatic labeling of duplicate records
- Clustering of USHMM person records
- Uses semantic database. Eg Maria Izsak in ONTO GraphDB vs USHMM HSV

Source: <http://data.ehri-project.eu/ushmm/person/4926452>

	subject	predicate	object
	subject	predicate	object
1	ushmm_person:4926452	onto:daitchMokotoffEncoding	"694500"^^xsd:string
2	ushmm_person:4926452	onto:firstNameDM	"MR"^^xsd:string
3	ushmm_person:4926452	onto:gender	onto_ushmm:4926452-gender-LinearClass
4	ushmm_person:4926452	onto:gender	onto_ushmm:4926452-gender-RuleBased
5	ushmm_person:4926452	onto:lastNameDM	"ASSK"^^xsd:string
6	ushmm_person:4926452	onto:nameDM	"MR ASSK"^^xsd:string
7	ushmm_person:4926452	onto:normalizedFirstName	"Maria"^^xsd:string
8	ushmm_person:4926452	onto:normalizedLastName	"Izsak"^^xsd:string
9	ushmm_person:4926452	onto:normalizedName	"Maria Izsak"^^xsd:string

**Holocaust Survivors and Victims Database** MEMORIAL MUSEUM 2018

**MARIA IZSAK**

Date of Birth: 16 Nov 1929  
Place of Birth: Lörinczfalu  
Occupation: Schneiderin  
Persecution Category: • Politische Ungeladen / Jüdinnen [Political and ideological targets]  
• Politische Ungeladen / Jüdinnen [Jew]

Prisoner Number: 63181

**Search for Names** NEW SEARCH

**Search for Lists**

**Register in the Survivor Registry**

**What's New**



# Names Normalization

- 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.

# ML Model for Person Gender

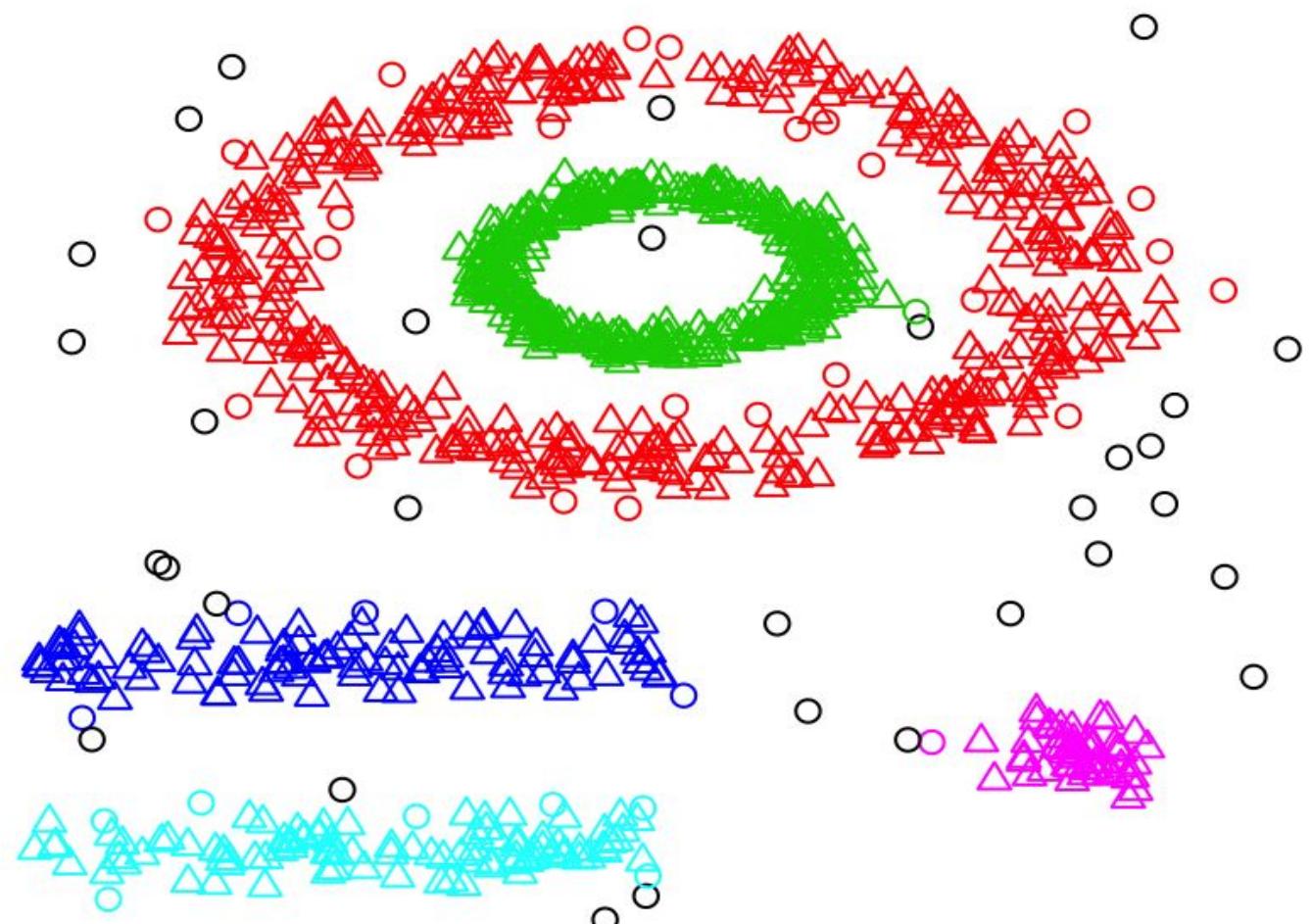
- Features to train ML model for automatic labeling of person gender
  - Name suffixes with various length (1, 2 and 3)
  - First name
  - First and last name
  - Nationality
- Rules
  - If there is another person with the same first name and known gender assign the same gender to the person with unknown gender
  - Same for the whole name
  - If there are multiple persons with different genders do not assign a gender

# Statistical Model for Person Matching

- Lexical similarities of names (person first and last name, mother name)
  - Jaro-Winkler Similarity
  - Levenshtein Similarity
  - Beider Morse Phonetic Codes Matching
- Birth dates similarity (approximate match)
- Birth place similarity: 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

# Person Clustering

- Grouping objects so those in a group are more similar than those in other groups
- Each cluster represents a possibly unique person, comprising 1 or many person records
- Example cluster:
  - VÁCLAV ŽIŽKA, born 20 Jan 1903
  - VÁCLAV ŽIŽKA, born 20 Jan 1903
  - VÁCLAV RŮŽIČKA, born 28 Jan 1903



# EAD Archival Descriptions

- Make EAD aggregation more sustainable
  - Convert from various formats (XML, tabular, JSON) to standard EAD
  - XML Validation: schema (XSD) and extra rules (Schematron)
  - Preview as HTML, show errors integrated in the preview
  - Enable transport and incremental update (synchronization)
  - Provide "self-service" functionality so archival institutions can initiate the process and validate results
- Info gathering, process
  - See e.g. USHMM info (google doc) Table of Contents on the right

- 1 Introduction
  - 1.1 Background
  - 1.2 Links
- 2 Collection Holding Institute
  - 2.1 CHI project manager
  - 2.2 CHI collection specialist
  - 2.3 CHI technical contact
- 3 EHRI contacts
  - 3.1 EHRI wp9 (content) contact
  - 3.2 EHRI wp10 (technical) contact
- 4 Metadata Evaluation
  - 4.1 General remarks
    - 4.1.1 Validation errors
    - 4.1.2 Identifiers
    - 4.1.3 Language
    - 4.1.4 Record Counts
    - 4.1.5 Field Analysis
    - 4.1.6 Hierarchy Analysis
    - 4.1.7 Links to Finding Aid and Digital Material
    - 4.1.8 Links to Parallel Descriptions
  - 4.2 Analysis of EHRI1 Export
    - 4.2.1 Field Occurrences
  - 4.3 Publishing/Transport Protocol, Sync
    - 4.3.1 datetimemodified Distribution
- 5 Discussion About Parent Items
  - 5.1 Virtual meeting 2017-05-15 (Linda)
  - 5.2 2017-06-06 (Vladimir, Boyan)
- 6 Selection Evaluation
- 7 Thesaurus Use

# EAD & Ingestion Task Tracking

The screenshot shows a Trello board titled "ehri2: IT-ICAR Selection&Import". The board is set to "Team Visible" and has the "Trello" logo at the top right. The interface includes a "Boards" tab, a search bar, and a "Show Menu" option.

The board has the following columns:

- todo by WP9**: Contains cards for "decide on selection" and "Install PMH harvester and collect the selected collections - T10.2".
- todo by WP10**: Contains cards for "advise the CHI on mapping tools" and "advise the CHI on publishing tools".
- doing**: Contains a card for "evaluate sample file".
- waiting for CHI**: Contains a card for "Add a card...".
- done**: Contains cards for "create form for this CHI", "enter contact info to form", "create b2drop account", "set up B2Drop for CHI", "request sample files from CHI", "send sample file to WP10", and "decide approach to importing".

Below the columns, there are several cards representing different projects or tasks, each with a yellow star icon:

- Import workflow default (EHRI)
- ehri2: SK-HDC selection&import (EHRI)
- ehri2: NL-NIOD selection&import (EHRI)
- ehri2: IT-CDEC selection&import (EHRI)
- ehri2: GR-JMG selection&import (EHRI)
- ehri2: CZ-JMP selection&import (EHRI)
- ehri2: BE-KAZDOSSIN selection&import (EHRI)
- ehri2: BE-CEGESOMA selection&import (EHRI)
- ehri2: DE-BARCH selection&import (EHRI)
- ehri2: IT-ASCR selection&import (EHRI)
- ehri2: US-USHMM selection&import (EHRI)
- euBusinessGraph Board (EHRI)
- EHRI2 wp13 (EHRI)
- ehri2: IT-ICAR Selection&Import (EHRI)
- ehri2: HU-HJA selection&import (EHRI)
- ehri2: LT-VGSJM selection&import (EHRI)
- ehri2: RO-INShR-EW selection&import (EHRI)
- WP10 workshops (EHRI)
- ehri2: IL-GFM selection&import (EHRI)
- ehri2: IL-Yad\_Vashem selection&import (EHRI)

# EAD Convertor



The logo for the European Holocaust Research Infrastructure (EHRI) features a dark purple square with a white stylized letter 'E' on the left and a white 'H' on the right. Below the letters, the text 'EUROPEAN HOLOCAUST RESEARCH INFRASTRUCTURE' is written in a smaller, white, sans-serif font.

**EAD CONVERSION TOOL**

DOCUMENTATION

VIEW GOOGLE SPREADSHEET

US-USHMM-mapping-config

target-path	target-node	source-node	value
/	ead	//doc	
/ead/	eadheader	.	
/ead/eadheader/	profiledesc	.	
/ead/eadheader/profiledesc/	creation		"This EAD is created by EHRI on ", <date>
/ead/	archdesc	.	
/ead/archdesc/	did	.	
/ead/archdesc/did/	unitid	if (.str[@name="accession_number"]/text() != .str[@name="id"]/text()) then .str[@name="id" text()]	attribute label ("accession_number"), te
/ead/archdesc/did/	unitid		attribute label ("former_accession_num
/ead/archdesc/did/	unitid		attribute label ("recordgroup_number")
/ead/archdesc/did/	unititle		attribute label ("subtitle"), text()
/ead/archdesc/did/	unititle		attribute label ("alternative"), text()
/ead/archdesc/rirt/	unititle		

Select Local Mapping

Select local mapping file

PREVIOUS STEP

NEXT STEP

- Universal conversion configured with a Google Sheet
- Specific conversions also supported

# EAD Convertor Configuration

- Analysis & conceptual mapping

field	oral-history	%	docs	%	kind	map to	agreed	mapping notes
accession_number	14013	92.7%	9559	99.4%	sng	did/unitid with @label="accession_number"	1	
accession_number_addl	579	3.8%	187	1.9%	arr	did/unitid with @label="former_accession_number"	1	previously used ids
acq_credit	9156	60.6%	3345	34.8%	arr	acqinfo	1	
arrangement	3	0.0%	2454	25.5%	sng arr	arrangement	1	
assoc_parent_title	12539	82.9%	1091	11.3%	sng arr	did/unittitle @label="alternative"	1	only use when different from collection_name

- Physical Mapping (XPaths everywhere)

target-path	target-node	source-node	value
/	ead	//doc	
/ead/eadheader/	eadid	./str[@name="id"]	text()
/ead/archdesc/did/	unitid	if (./str[@name="accession_number"]/text() != ./str[@name="id"]/text()) then ./str[@name="id"] else ()	text()
/ead/archdesc/did/	unitid	./str[@name="accession_number"]	attribute label {"accession_number"}, text()

# Conclusion

- We presented 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.

# Thank you!

Question time