Semantic Archive Integration for Holocaust Research: the EHRI Research Infrastructure

Vladimir Alexiev & Ivelina Nikolova
Ontotext Corp
DSDH 2017, Venice, 30 June 2017

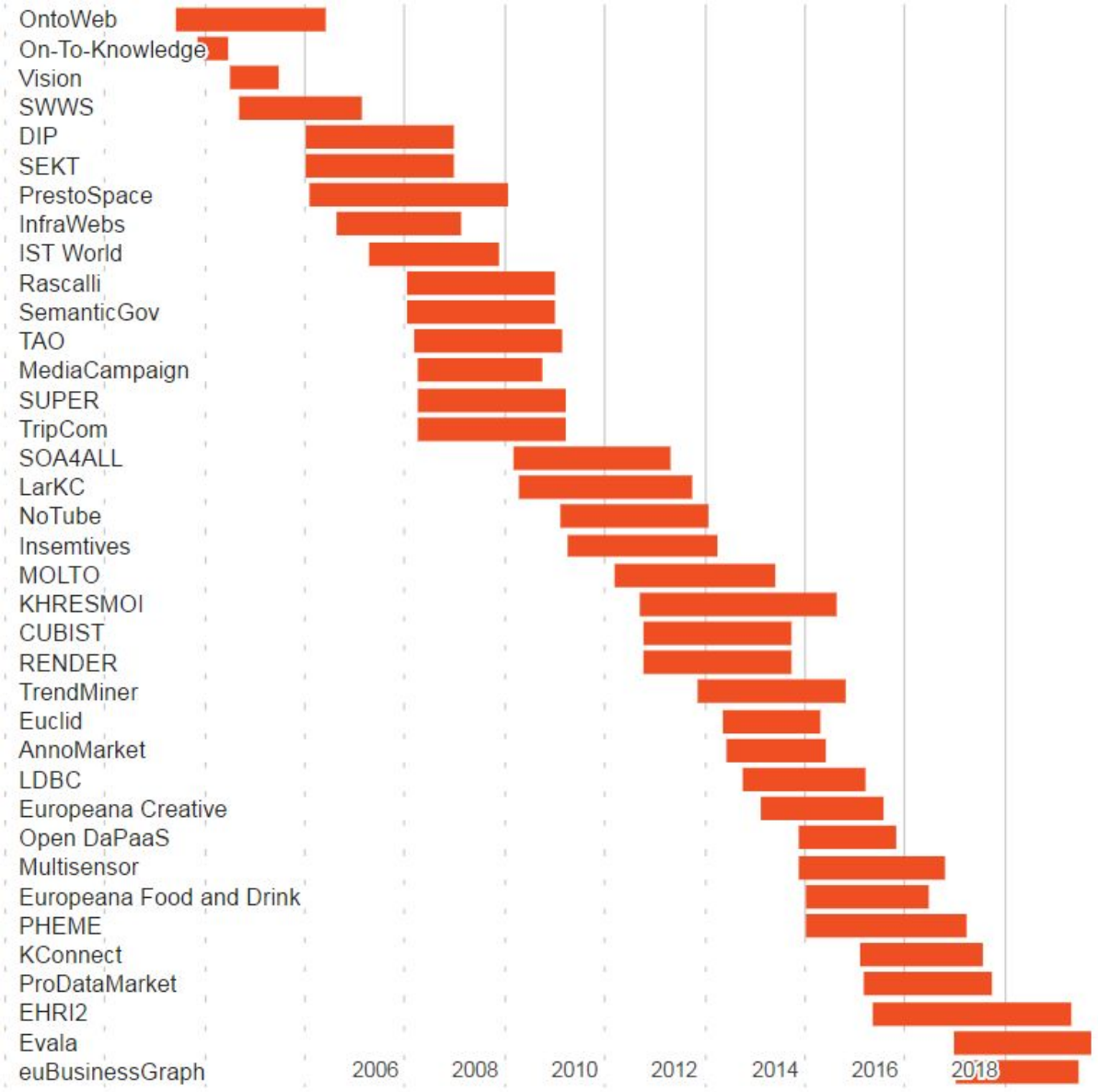
Ontotext

- 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)

ontotext

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

ontotext

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

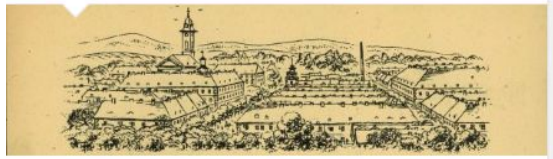
- Names and Networks: **Chances of Survival during the Holocaust**. Most Jews needed the support of other people to survive. Persons found aiders inside their personal and group networks. Investigate the networks in which European Jews operated during their persecution in the Second World War, and improve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

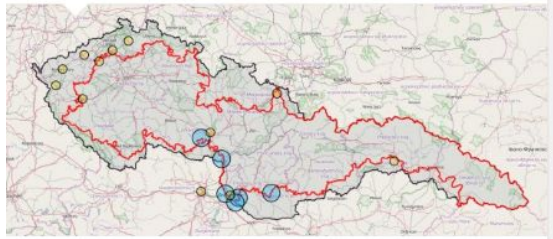
Testimony of Valerie Straussová

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 –…

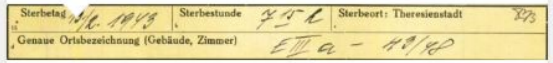
Continue Reading →

2016/03/11

Reports from the No Man's Land



Death Certificate of Gabriel Frankl from the Terezín Ghetto



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

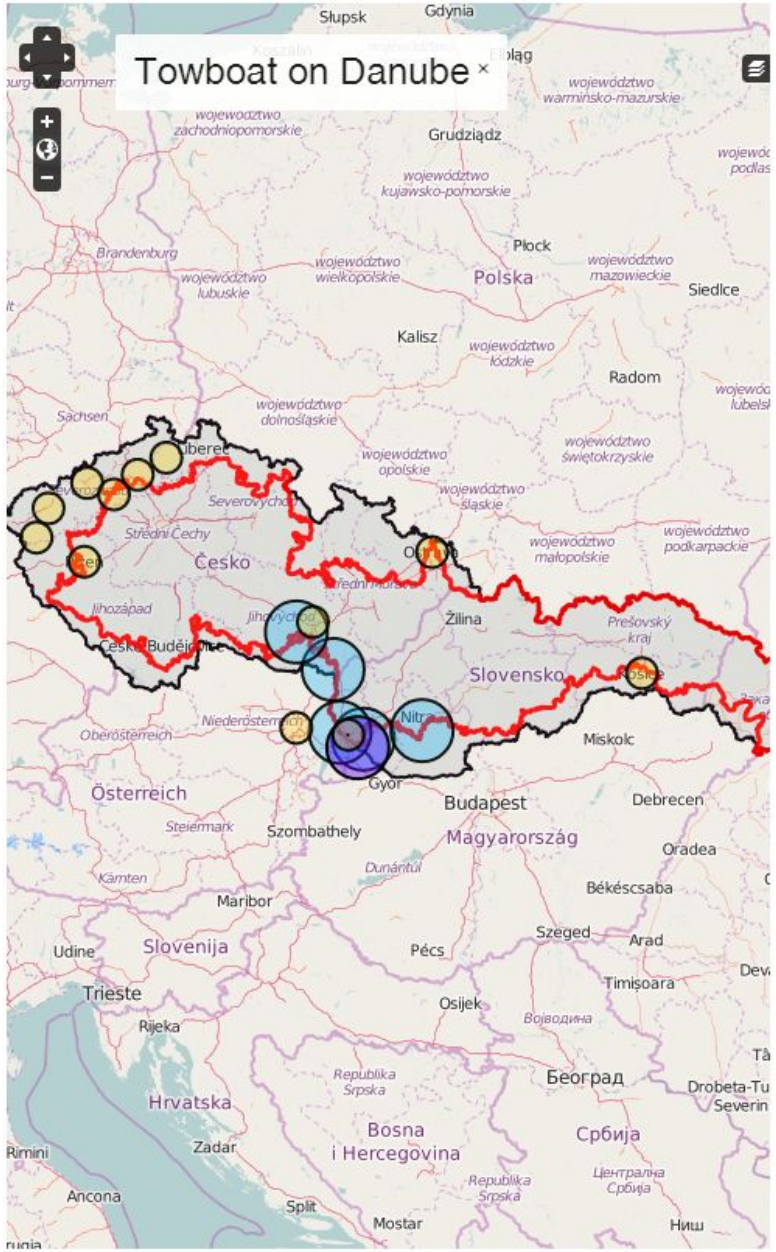
Making digital collections and archival material can be full of challenges. As more and more material becomes accessible online, it also means that archivists and historians have to think up new ways to help users get the most out of…

Continue Reading →

- Provides inspiration for:
  - Research questions
  - Research techniques
- Our task:
  - Enable such tools at scale
  - Help researchers use them
- Soliciting other interesting examples!

ontotext

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

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

ontotext

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%			

ontotext

Controlled Access Point Problems

- 23 ways of spelling **Łódź** across instiutions.
  - 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,גטו,Poland (note: "גטו" means Ghetto)
Lodž
Lodž jc-places-place-iti-48 cvocConcept
Lodž terezin-places-place-iti-48 cvocConcept

ontotext

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	# **corporateAccess**: 8th SS Cavalry Division Florian Geyer

- 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

ontotext

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

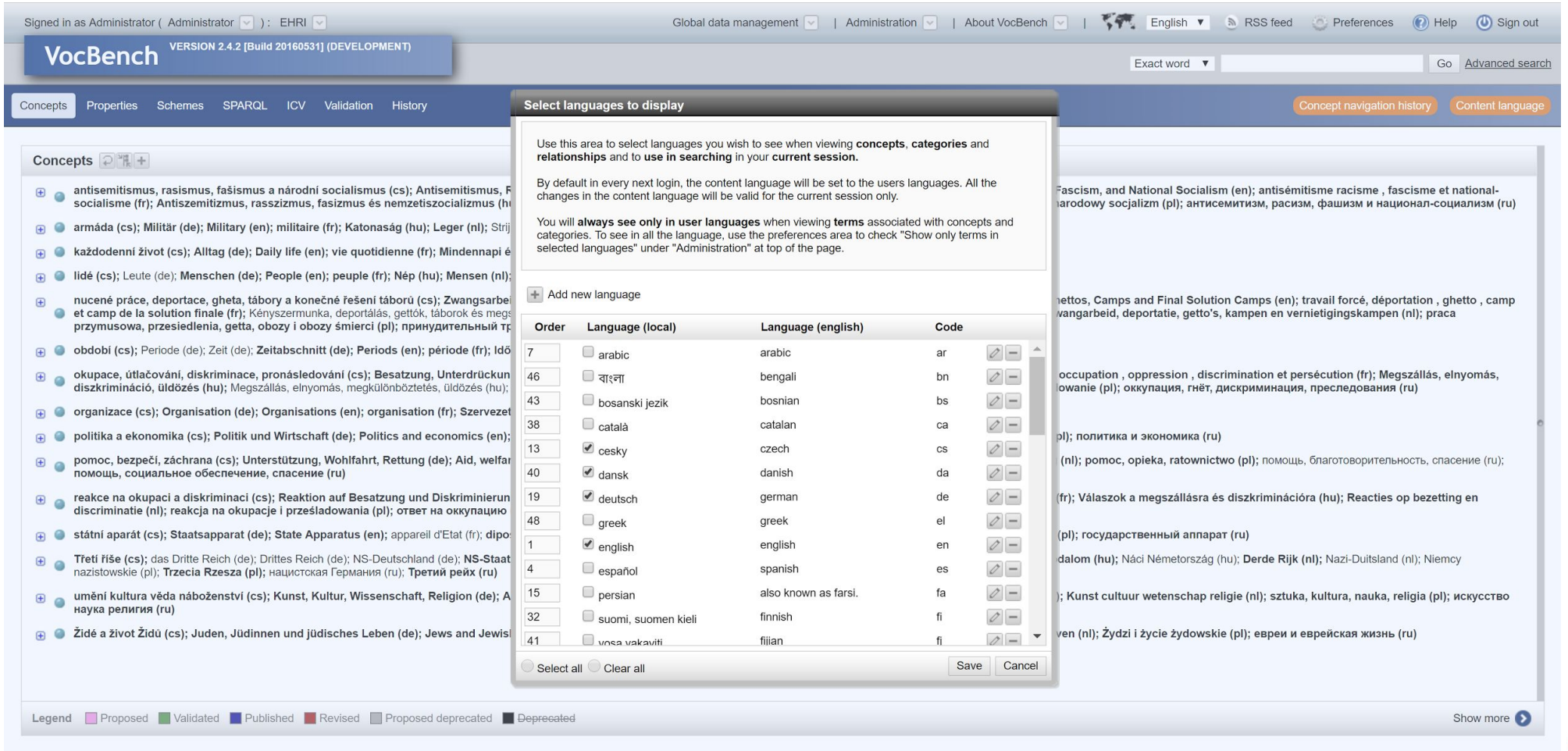
ontotext

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

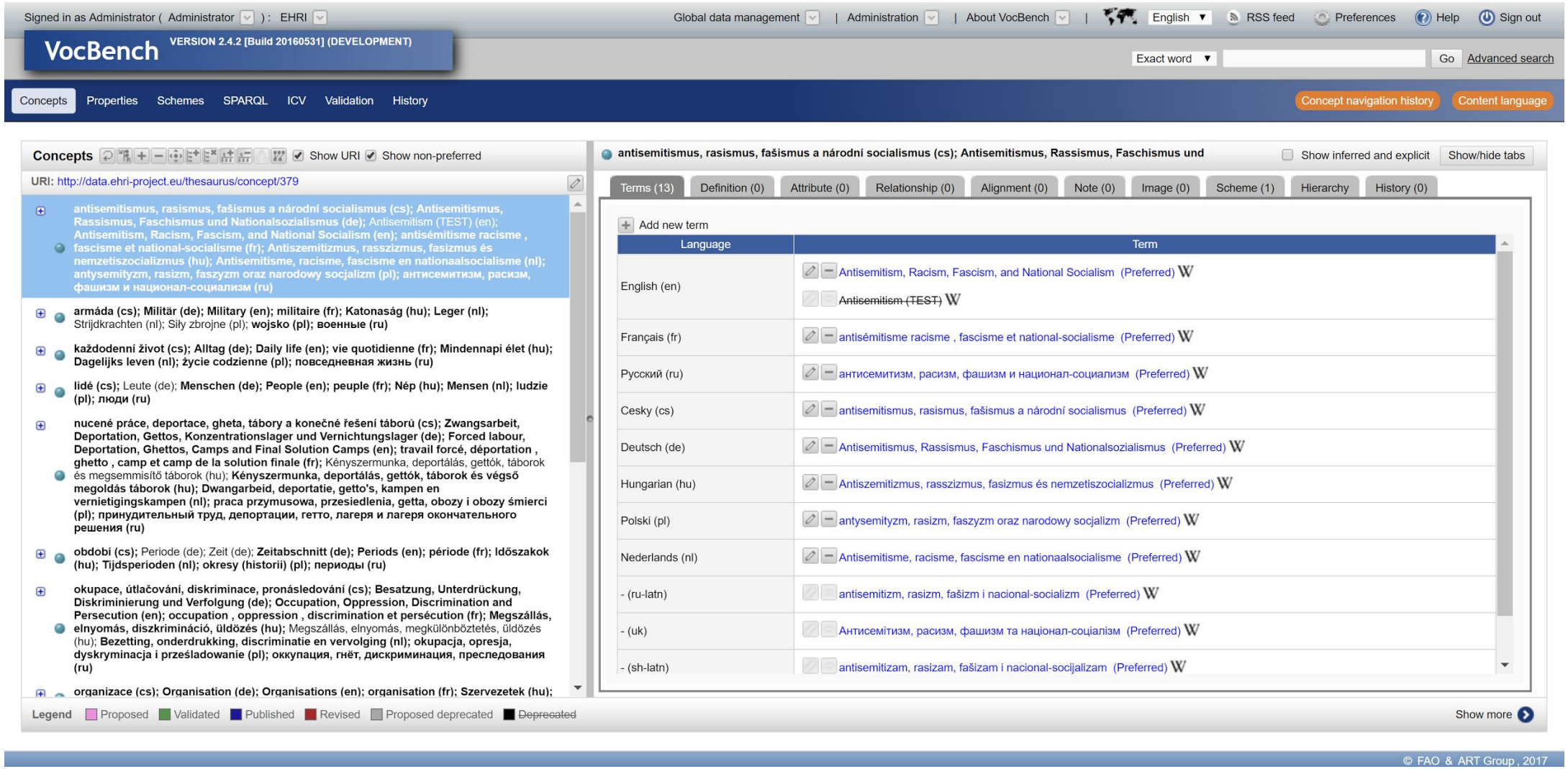
- VocBench: Best open source TMS
  - Runs directly over semantic data (ONTO GraphDB); partnership with developer (UniRoma2)
  - Used by EC OPOCE (EuroVoc), UN FAO (AgroVoc)
  - E.g. multilingual labels



ontotext

Subjects/Concepts: VocBench

- Concept information



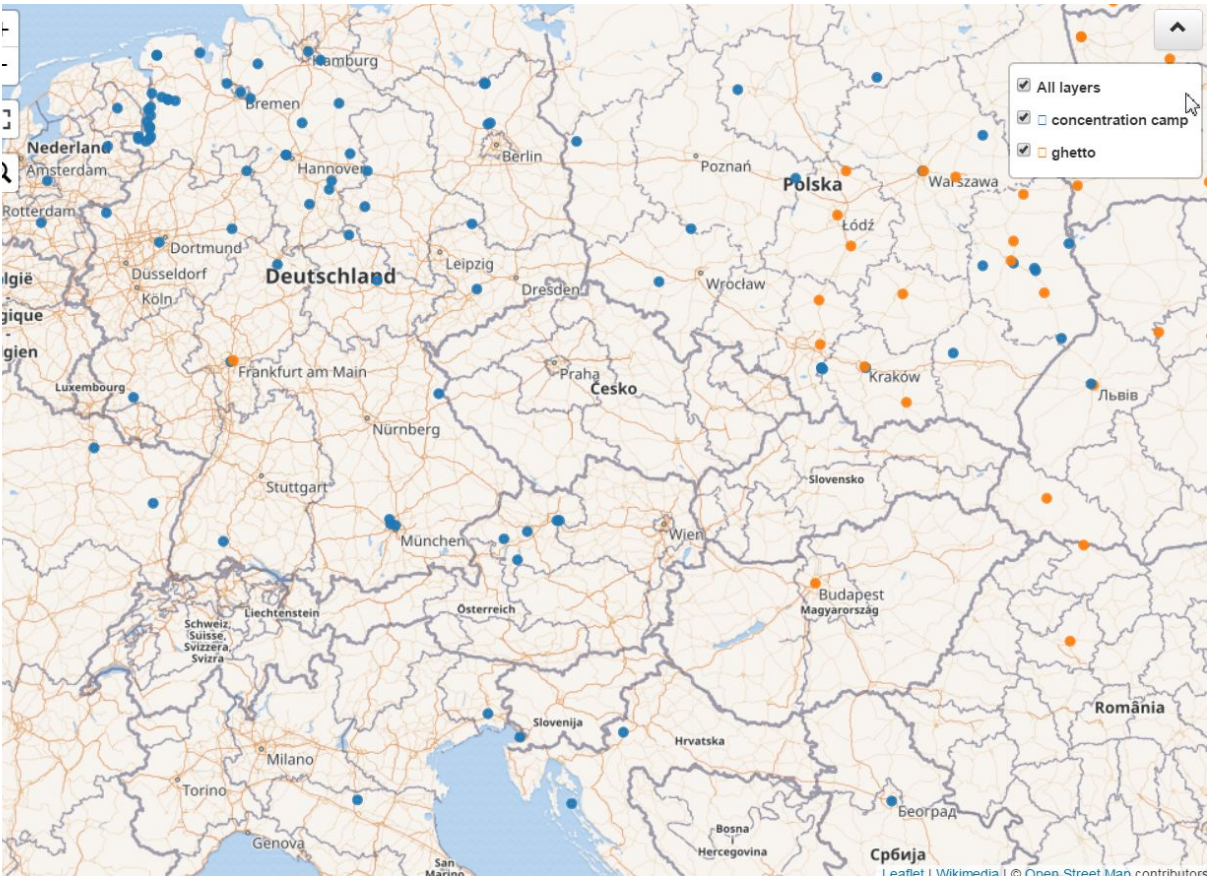
ontotext

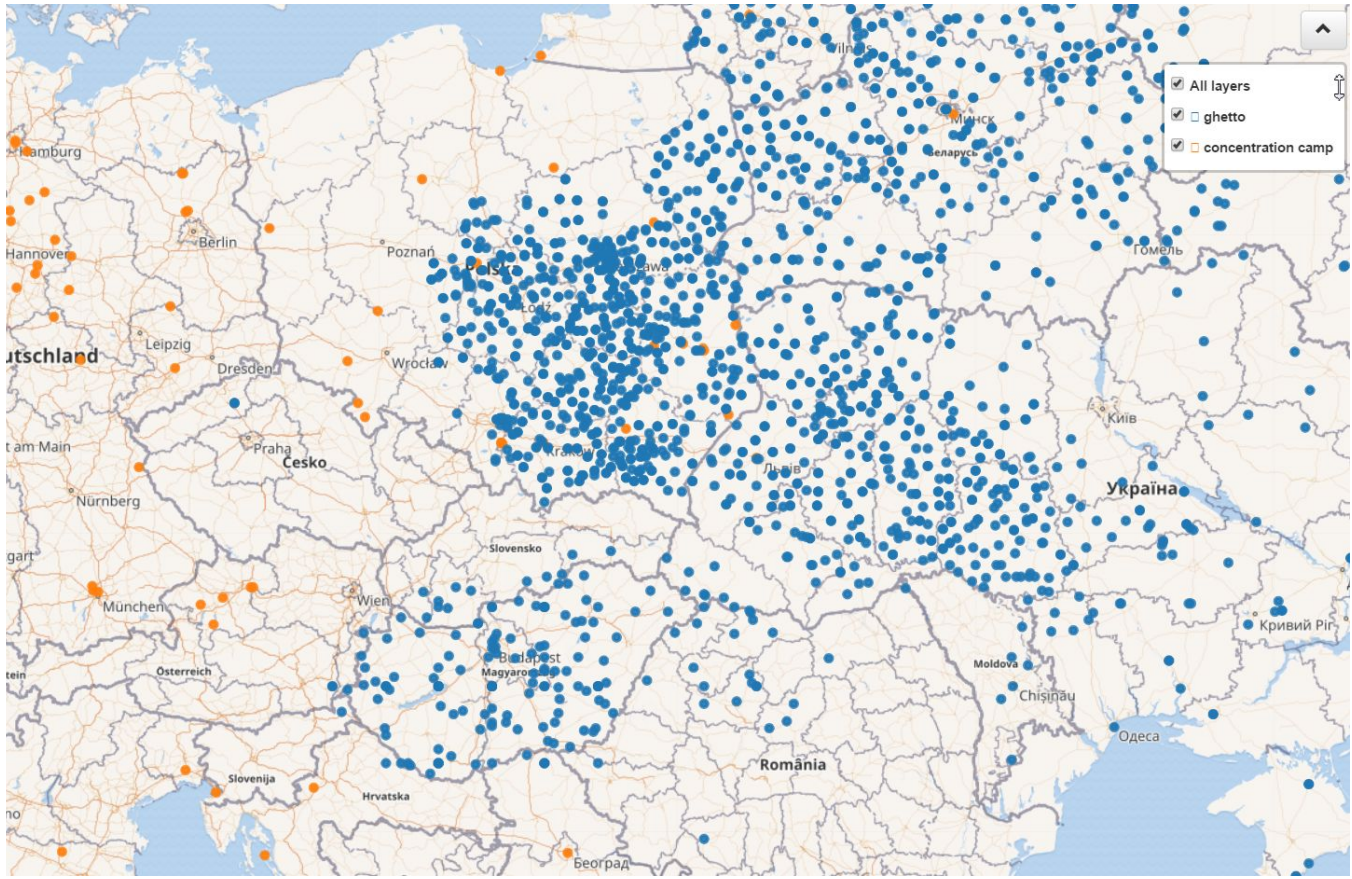
Ghettos & Camps

- 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 Trostinets, Maly Trastsianiets, 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

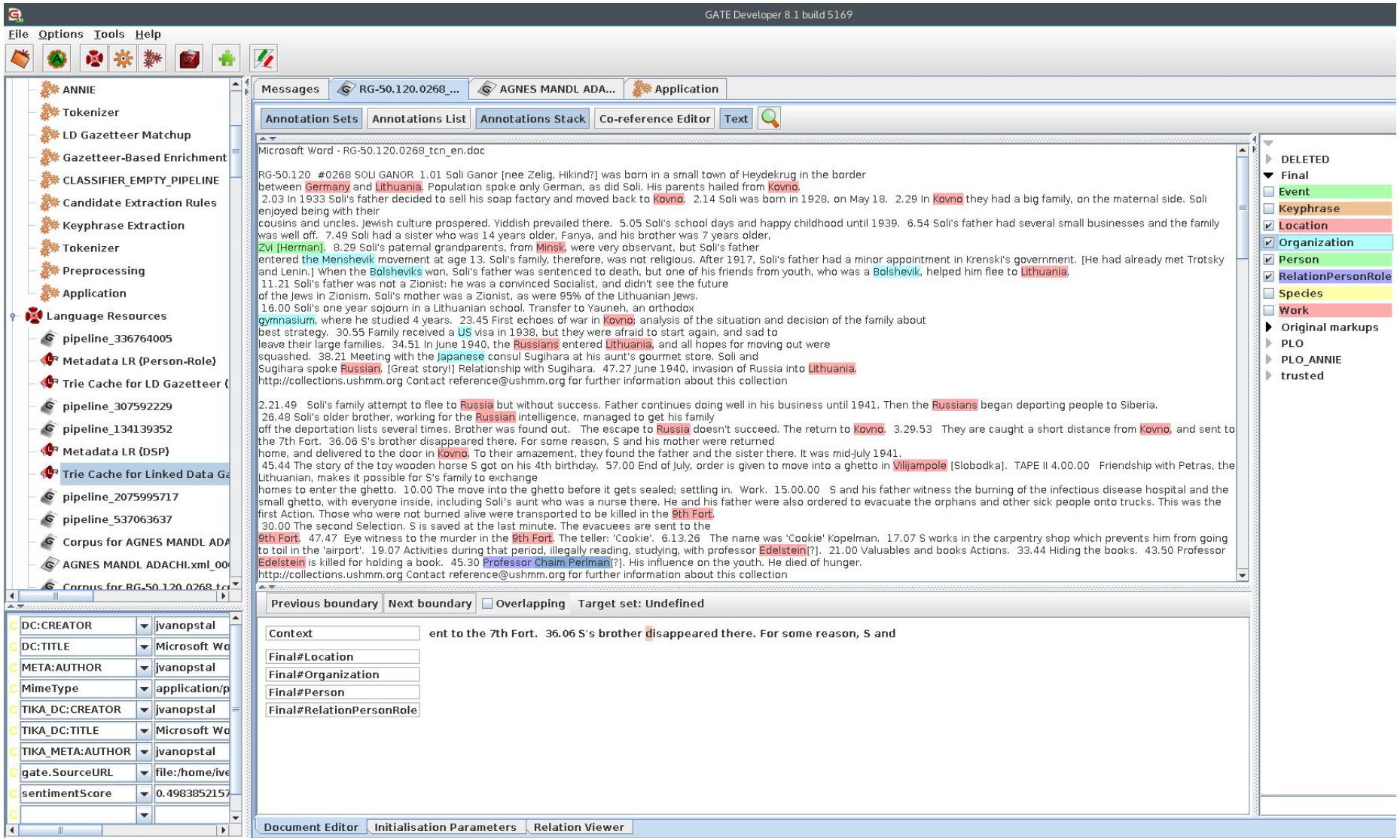
- EHRI2 decided to coreference to Wikidata
  - So we can get lots more labels, coordinates, place/location…
  - And international collaboration! E.g. Wikidata.GLAM Facebook Group
  - Camps and Ghettos on Wikidata: before (15 Mar 2017) and now (30 Jun 2017):





ontotext

Semantic Annotation (Text Enrichment)



- **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

ontotext

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

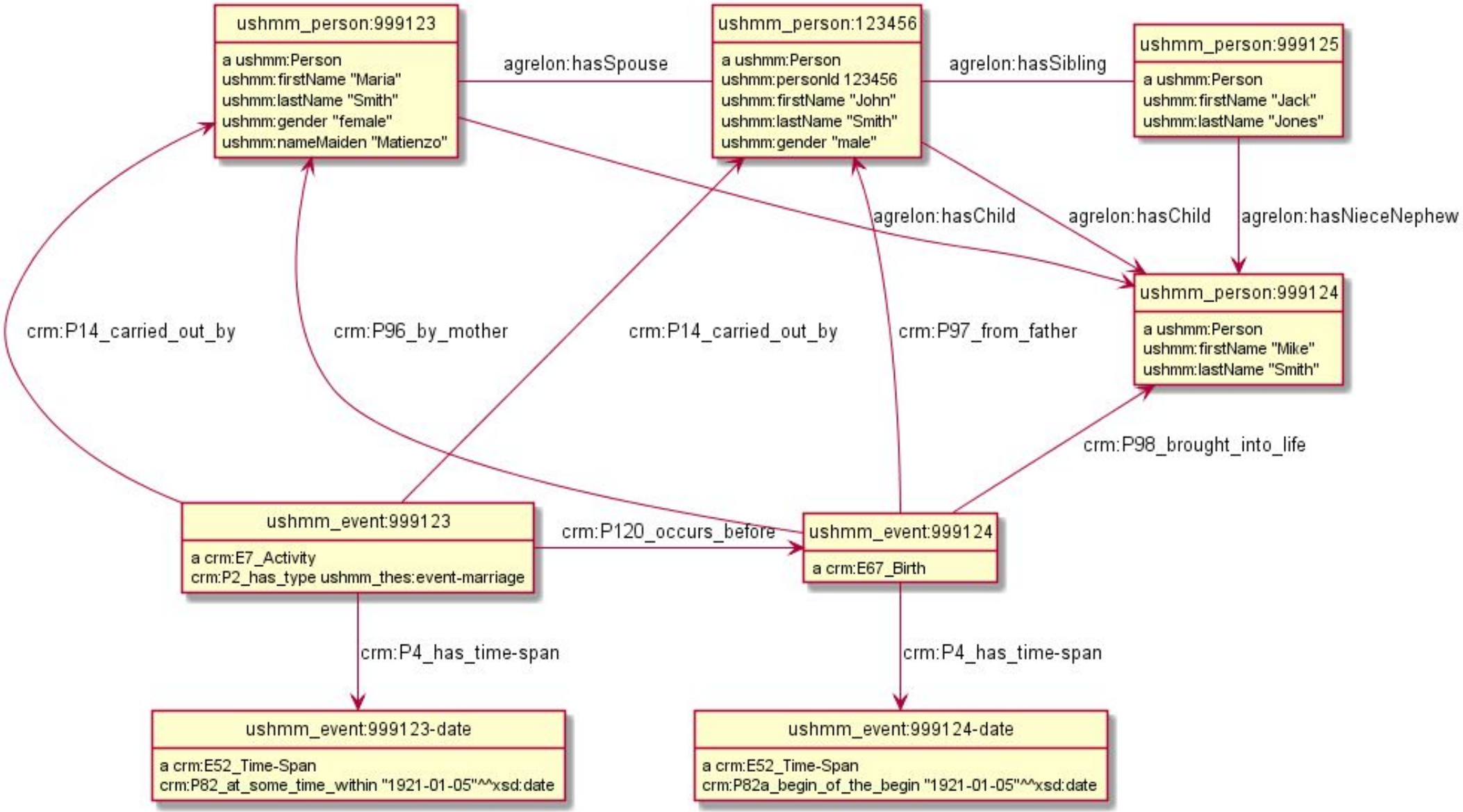
Jewish Social Networks

- 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

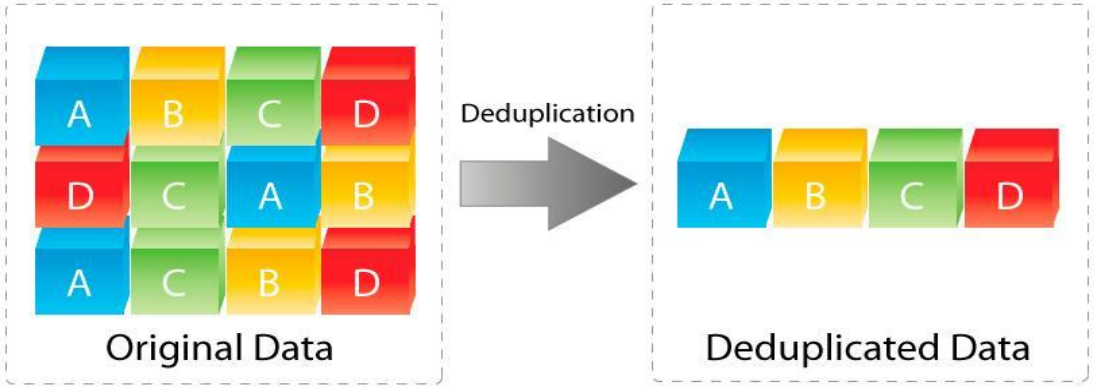
Reconstructing Families



ontotext

Person/Family Identification: Judgements

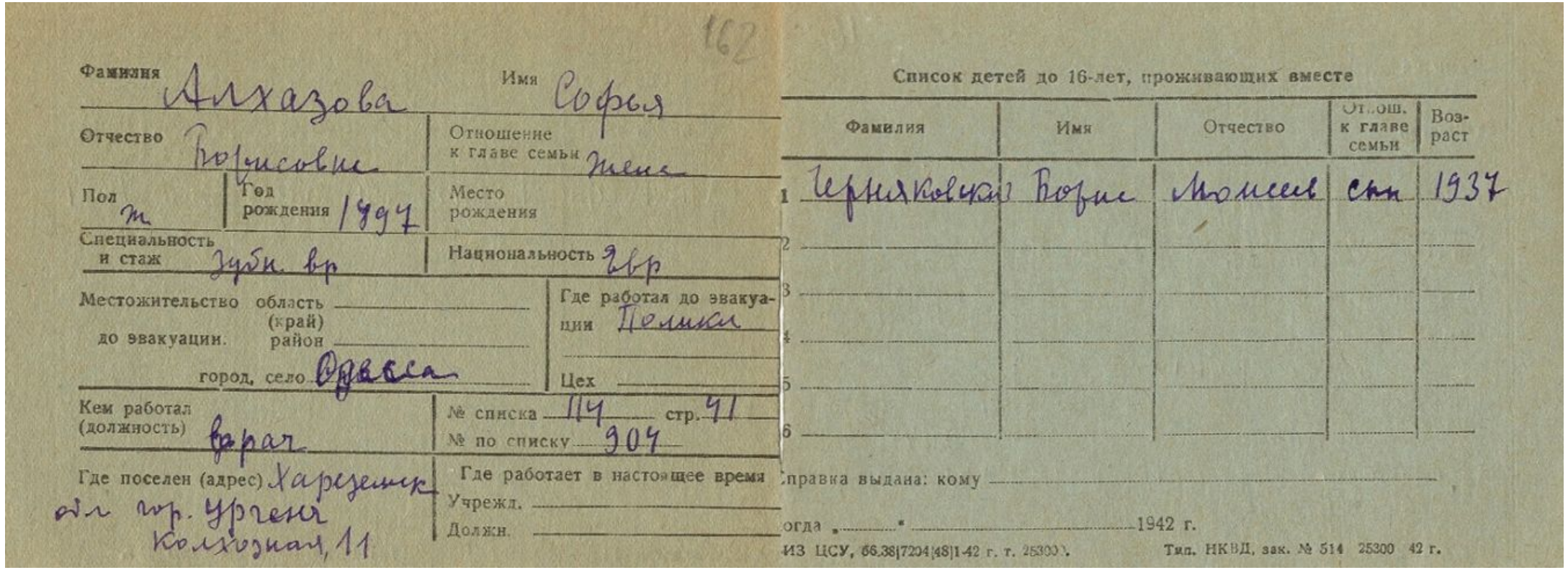
- Gold Standard with human judgement about possible matches
  - Using name search and historic reasoning
  - Difficult because of incomplete info
- Then we train a Machine Learning model for matching



- 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



Фамилия Алхазова Имя Софья
Отчество Борисовна Отношение к главе семьи жена
Пол ж Год рождения 1904 Место рождения
Специальность и стаж зубн. вр Национальность евр
Местожительство область (край) район до эвакуации. город, село Одесса
Где работал до эвакуации Поликл
Цех
Кем работал (должность) врач
№ списка 114 стр. 41
№ по списку 904
Где поселен (адрес) Харезмск обл ул. Ургенч Колхозная, 11
Где работает в настоящее время Учрежд. Должн.

Список детей до 16-лет, проживающих вместе

	Фамилия	Имя	Отчество	Отношение к главе семьи	Возраст
1	Черняковский	Борис	Моисеев	сын	1937
2					
3					
4					
5					
6					

Справка выдана: кому
Когда "..." 1942 г.
ИЗ ЦСУ, 66.38|7204|48|1-42 г. т. 25300. Тип. НКВД, зак. № 514 25300 42 г.

ontotext

Historical and Place Reasoning

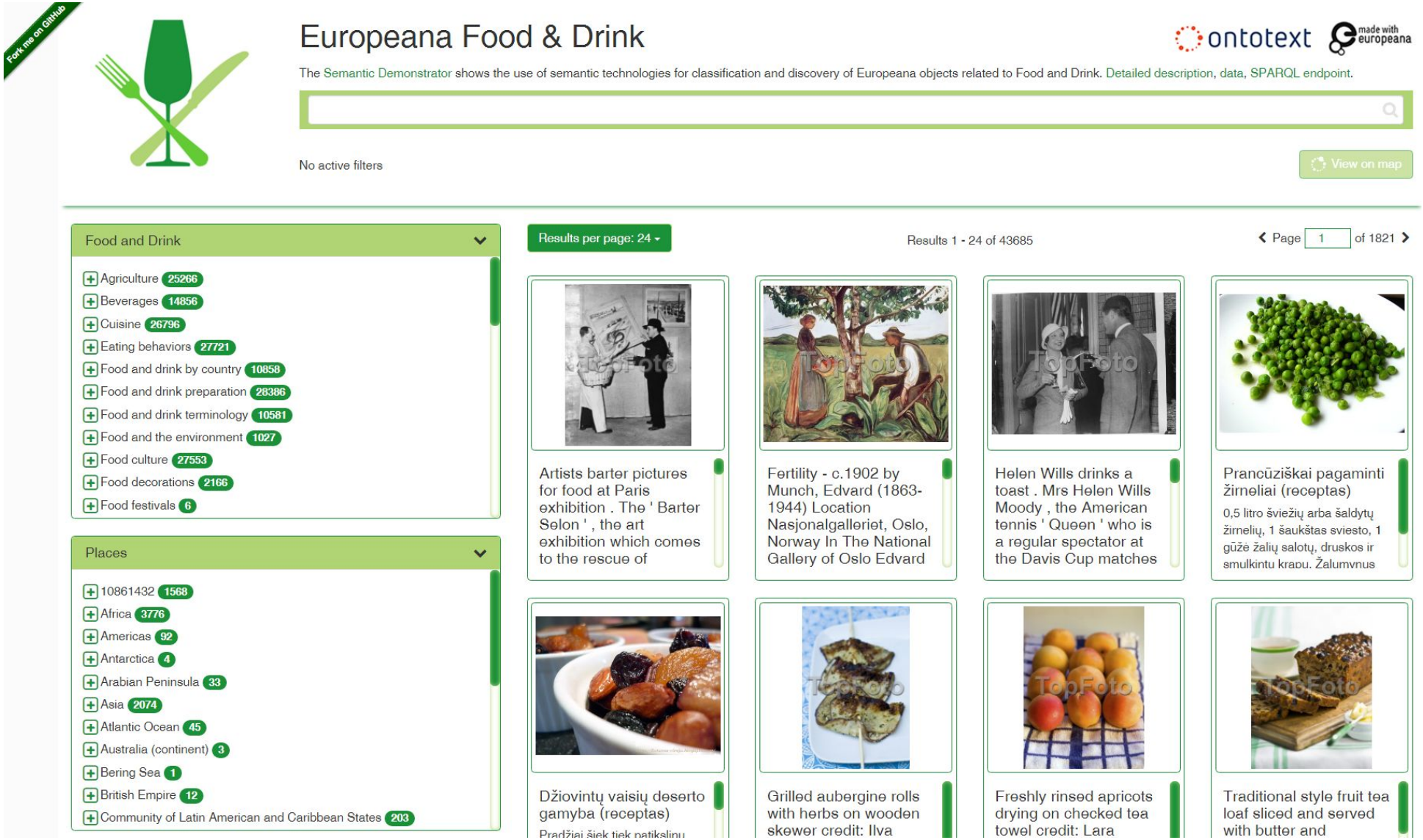
- After coreferencing to Geonames, we can use place hierarchy and proximity (based on coordinates) to make likely inferences.
- E.g. OH interview RG-50.661.0001:
  - "My Grandfather owned a bank in partnership with a cousin, Leibela Mandell"
  - "The family remained in Poland until the second world war"
  - 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



ontotext

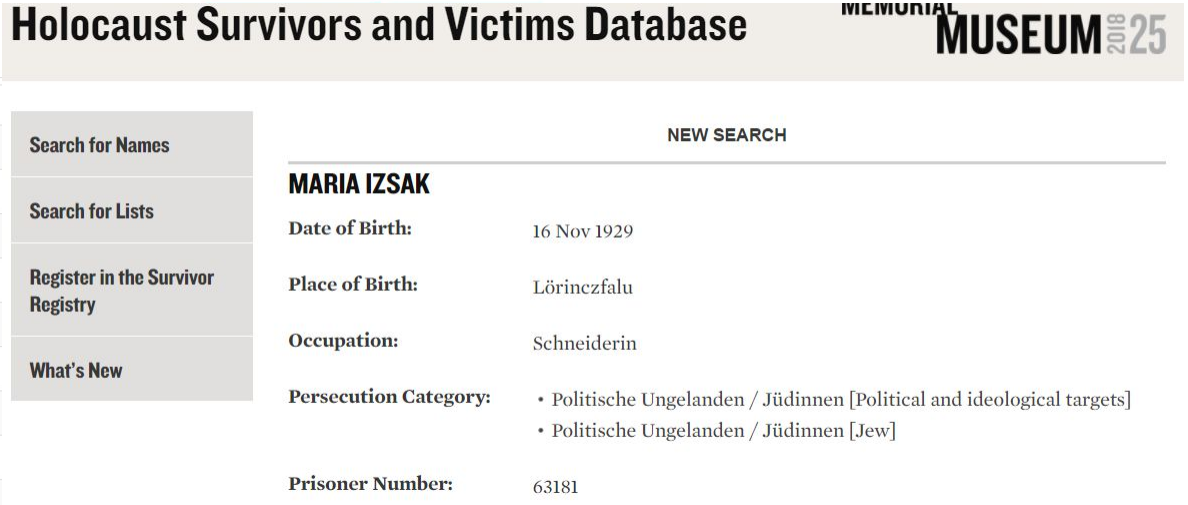
Person Deduplication

- Data preprocessing
  - Person Name Normalization, Daitch-Mokotoff encoding, Beider-Morse Phonetic (eg Maria Iszak -> 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 | context | all

	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



ontotext

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

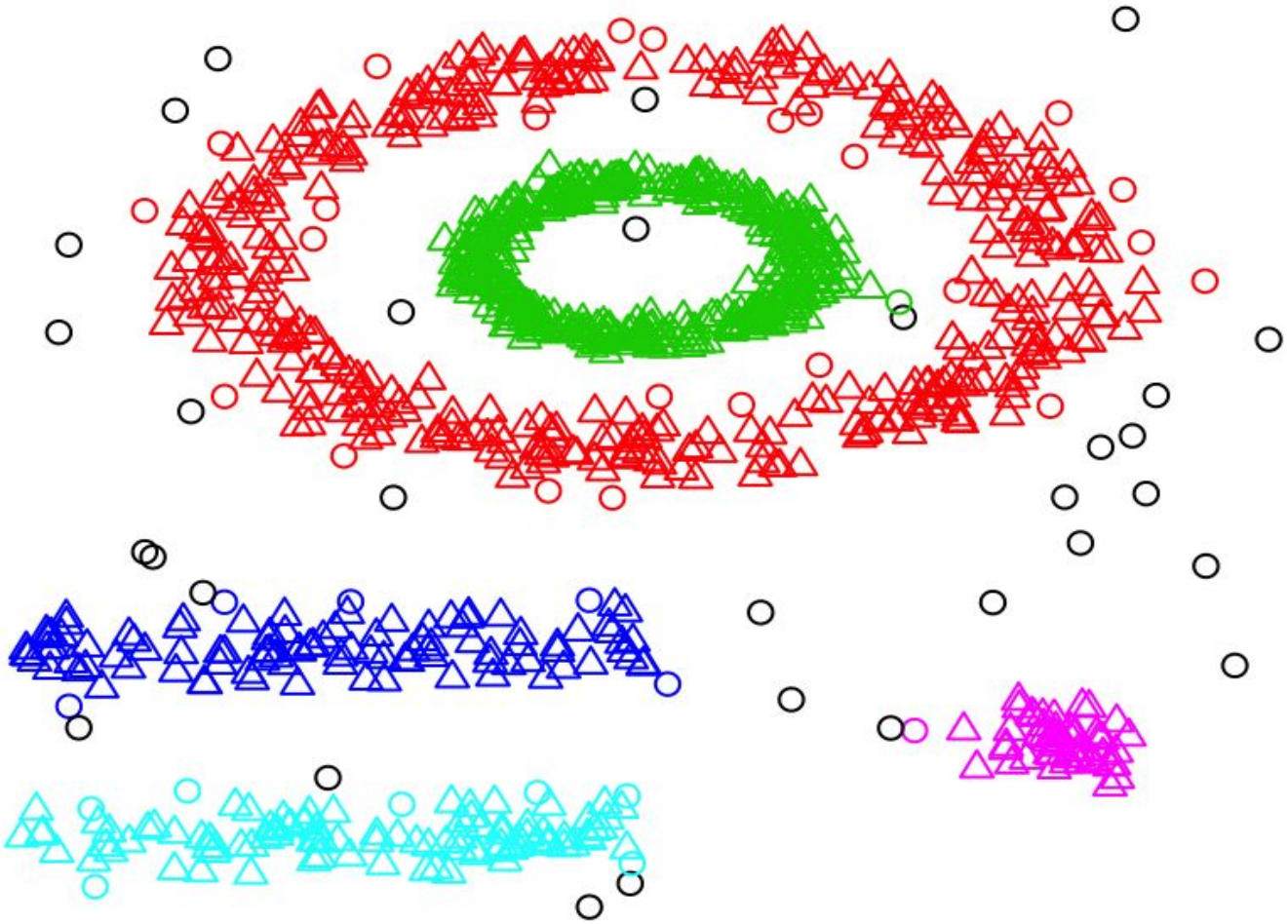
- Features to train ML model for automatic labeling of person gender
  - Name suffixes with various length (1, 2 and 3)
  - First name
  - First and last name
  - Nationality
- Rules
  - If there is another person with the same first name and known gender assign the same gender to the person with unknown 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

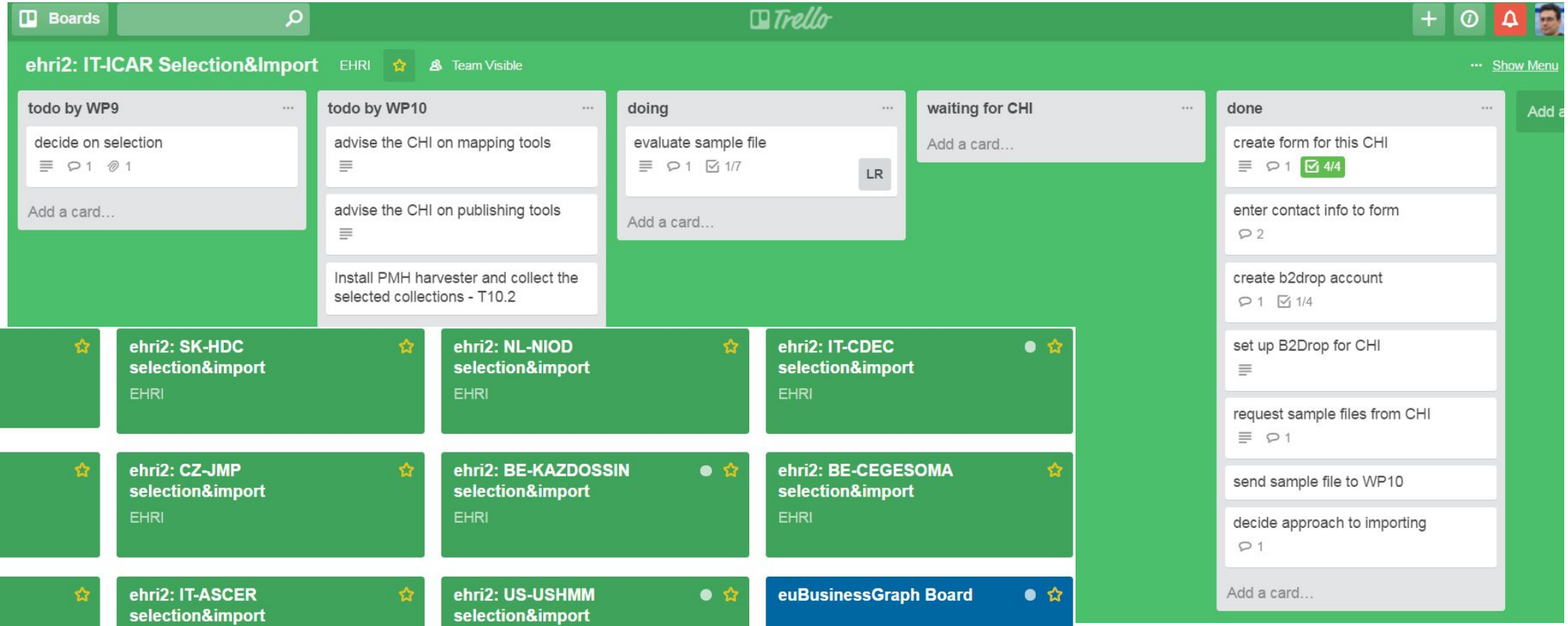
- Grouping objects so those in a group are more similar than those in other groups
- Each cluster represents a possibly unique person, comprising 1 or many person records
- Example cluster:
  - VÁCLAV ŽIŽKA, born 20 Jan 1903
  - VÁCLAV ŽIŽKA, born 20 Jan 1903
  - VÁCLAV RŮŽIČKA, born 28 Jan 1903

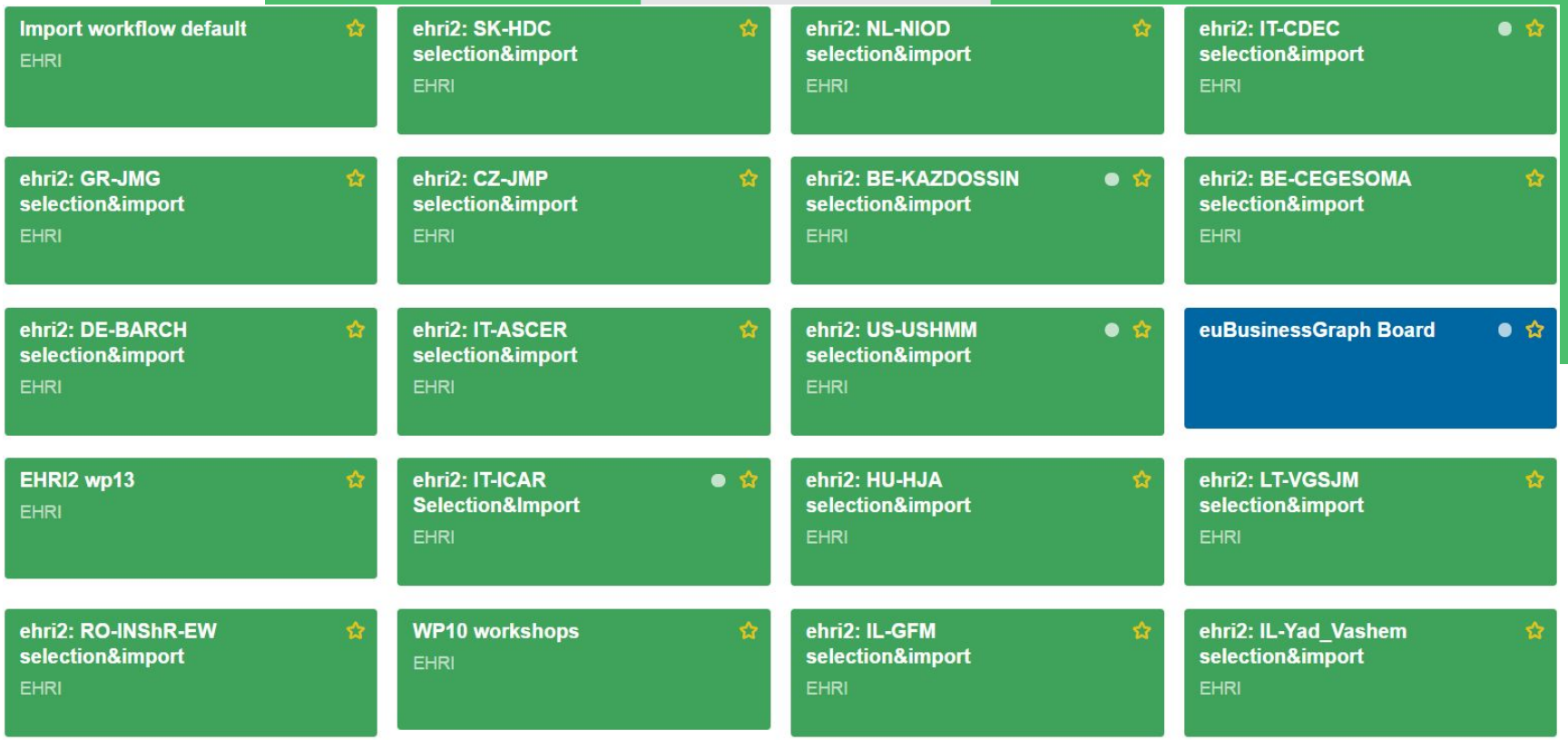


EAD Archival Descriptions

- 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

EAD & Ingestion Task Tracking

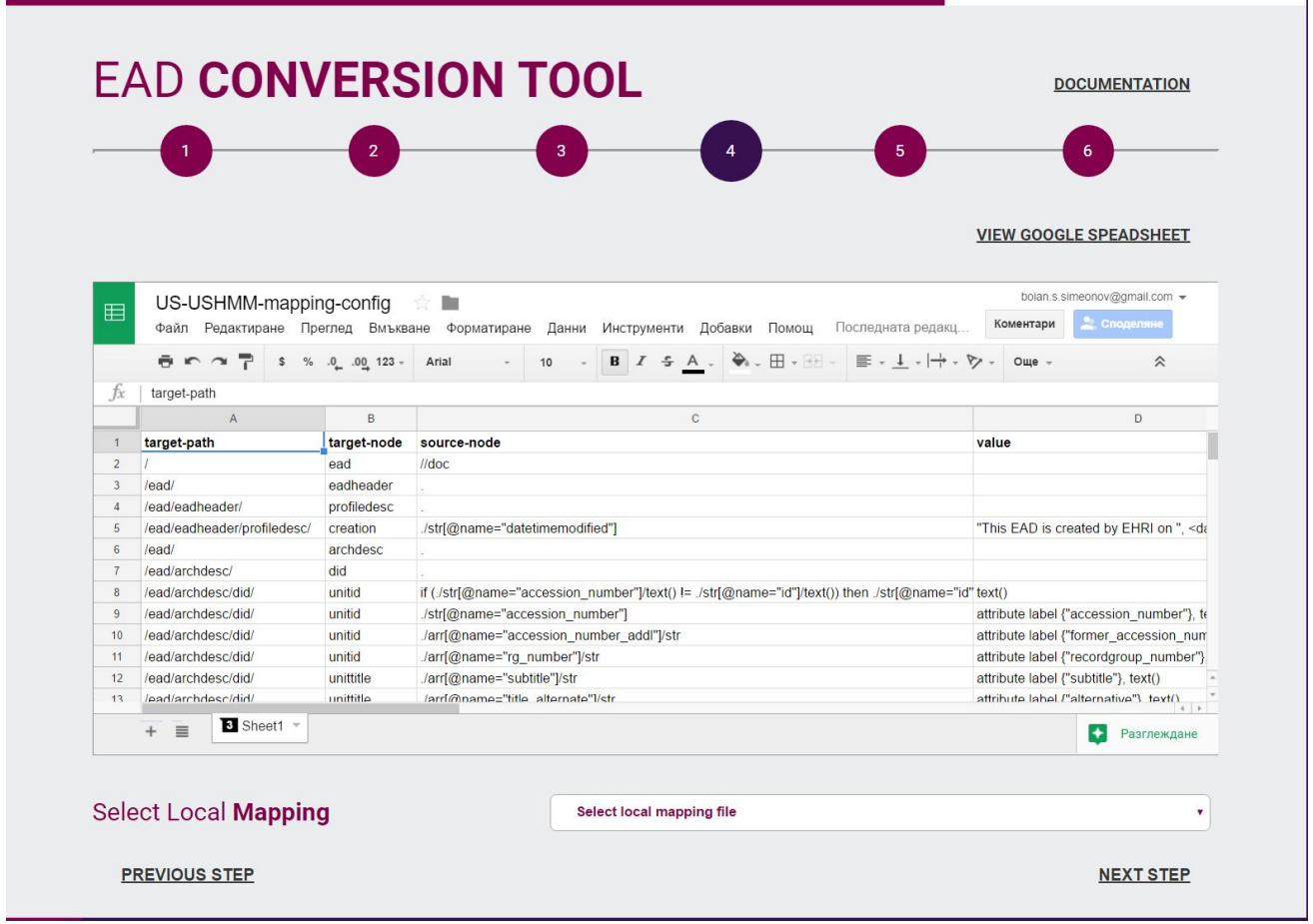




ontotext

EAD Convertor

EUROPEAN HOLOCAUST RESEARCH INFRASTRUCTURE



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

ontotext

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()

ontotext

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