EHRI Vocabularies and Linked Open Data: An Enrichment?

Annelies van Nispen

EHRI VOCABULARIES AND LINKED OPEN DATA: AN ENRICHMENT?

Annelies VAN NISPEN

Introduction

The European Holocaust Research Infrastructure (EHRI) started in October 2010 to build on a network that connects both people (Holocaust researchers, archivists, curators, librarians and digital humanists) and dispersed Holocaust source material and collections. EHRI's aim is making sources visible in a systematic way in order to counteract the fragmentation of the sources and to reveal interconnections. EHRI focuses on Archive and collection descriptions, which are available through the EHRI Portal[1]. EHRI is currently in its second phase and is on the ESFRI Roadmap[2] for a more sustainable future.

EHRI has developed a set of controlled vocabularies that serves both as a retrieval and cataloguing tool for the multilingual and highly heterogeneous data of the EHRI portal. These vocabularies were partly implemented in the first phase of the project. In the current phase of EHRI the vocabularies are in the process of quality improvement improve and enrich the existing terms, add new terms, disambiguate and remove the mistakes (deduplication, merging, adding multilingual labels, consistency checks, multiple parent relations, etc.) and increase their coverage. In the EHRI portal the subject terms are currently not available for the public, as they are used only for retrieval purposes.

EHRI vocabularies and Linked Open Data

EHRI's controlled vocabularies consist of the EHRI Thesaurus (subject terms), person and corporate body authorities and authority lists designed for the Holocaust knowledge domain as camps, ghettos, administrative districts and places. Some of the vocabularies of the project have already achieved a stable version, as, for example, the hierarchy of Administrative Divisions in the German Reich and Nazi occupied territories. The main goal of this project is to make the EHRI vocabularies efficient multilingual retrieval tools for the end users of the portal, and to serve as an efficient cataloguing and integration tool for newly ingested archival materials. These are the knowledge base for new developments as Named Entity Recognition (NER) and Linked Open Data (LOD).

---

[1] https://portal.ehri-project.eu.

[2] http://www.esfri.eu/esfri-roadmap.

Linked Open Data is a way of publishing structured data that allows metadata to be connected and enriched, so that different representations of the same content can be found, and links can be made between related resources. The strength of using Linked Open Data techniques and creating references to thesauri/controlled vocabularies is that it helps to connect and explain data. Single items in collections are remodeled into interlinked sources that reveal the bigger picture. EHRI started experiments with LOD to make use of these promising developments.

For the geographic locations the large geographical web database GeoNames[3] was used, for the two sets Camps and Ghettos Wikidata was used and for persons an experiment was conducted to match with VIAF.

Places and GeoNames

Historical events and historical resources have references to physical places. Some places still exist, some places have changed or disappeared. Users of digital data are interested in and search for named places, but marking and presenting places mentioned in resources can be complicated. EHRI is building a set of geographic locations connected to the Holocaust, by using GeoNames, the largest open geographical database on the web. In trying to connect both databases, the team encountered several difficulties. GeoNames is a database that is developed for the current geographic constellation, but historical locations are not a standard part of it. In the EHRI set there are many locations that are in countries/area's that from the Second World War and have geographical information from the war period. An extra complication is that many of the locations are in Eastern Europe where borders have shifted and places names have changed.

The Geo Names reconciliation had many issues, a few examples illustrate the difficulties:

- Historical place descriptions are not part of GeoNames (e.g., Altreich)
- Places listed in GeoNames but missing spelling variants (e.g., Babyn Iar)
- Typing errors or misspelling of location in the original data (e.g., Aushwitz instead of Auschwitz)
- A place name has more than one location (e.g., Berlin can be mapped to 176 different locations)
- Access points which are difficult to disambiguate without context (e.g., Bauer can be the German word for "peasant", a German family name, or a German town) and vice versa, where place names as Amsterdam get filtered out as person names
- Problem: Historical states, such as Yugoslavia or Czechoslovakia, are not properly linked to parents/children in the GeoNames data set

[3] https://www.geonames.org.

The result of the experiment with GeoNames reconciliation is that trying to find matches for Holocaust related geographic location has many issues. Although GeoNames is developing and many spelling variations and sometimes historical information is added to it, there are too many mismatches. Very strict manual quality control is required on the (semi-)automatic matching processes, which is labour-intensive work. The result of this experiment is that it could be useful in the future, but requires intensive manual quality control to avoid mismatches and incorrect information.

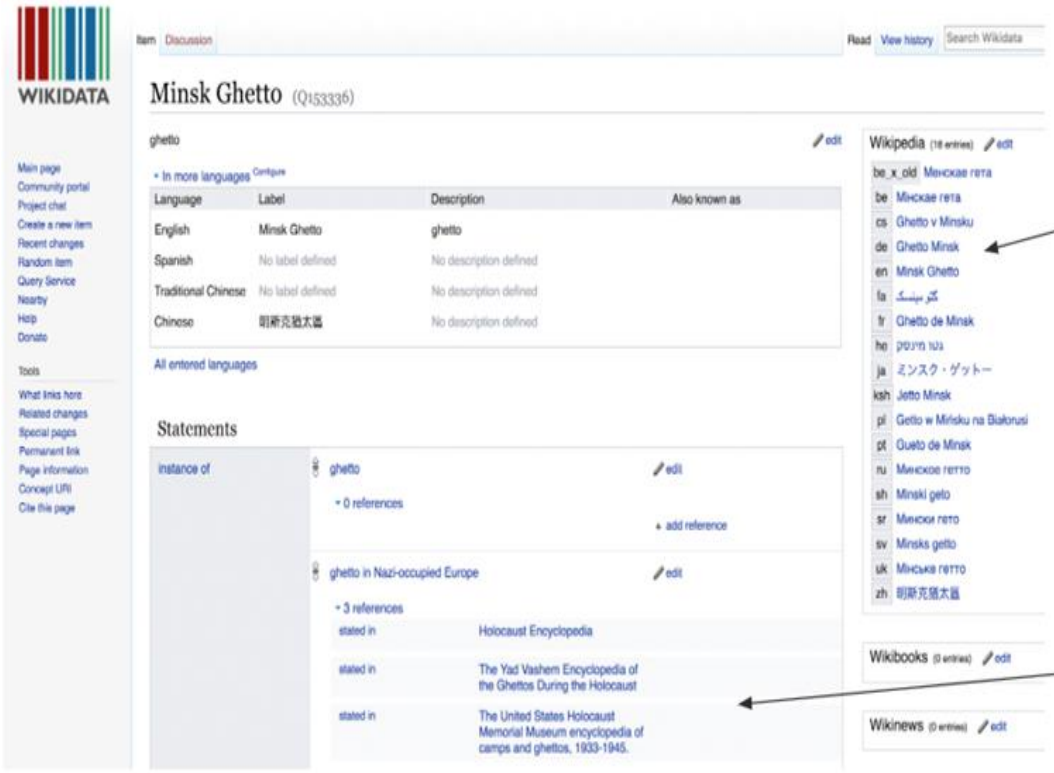
Ghettos, Concentration Camps and Wikidata

EHRI also works on Authority lists for Ghettos and Camps. These data are derived mainly from The Yad Vashem Encyclopedia of the Ghettos, The United States Holocaust Memorial Museum Encyclopedia of Camps and Ghettos and data from the Bundesarchiv. For these data we did an experiment to research if collaborating with Wikidata would enrich both Wikidata and EHRI.

Wikidata has proven to be an effective data integration platform to other multi-institutional initiatives. It has a flexible data model that can be extended for any purpose, can easily ingest data in various formats, and correlate the data. The large and diverse user base of Wikidata offers the benefits of continual crowdsourced data and makes EHRI more discoverable to Holocaust researchers.

We started with the data on ghettos from the EHRI portal. This data set contains a manageable 1,391 items, 80 of which already had identities in Wikidata. The intended results were to develop a workflow for importing data into Wikidata; establish lines of communication with the Wikidata community; and evaluate the usefulness of Wikidata as a data integration platform for EHRI. The data were imported in Wikidata and was enriched with the administrative territorial entity, alternative names or spellings, geographic coordinates and included dates of the ghetto (if available). This process was fairly straightforward given that all of the ghettos were pulled from reliable sources, including the EHRI portal, The Yad Vashem Encyclopedia of the Ghettos, and The United States Holocaust Memorial Museum Encyclopedia of Camps and Ghettos.

Both the Yad Vashem Encyclopedia of the Ghettos and The United States Holocaust Memorial Museum Encyclopedia of Camps and Ghettos include detailed information on the location of the ghettos, so these geographic coordinates were incorporated. These data were linked to EHRI, Yad Vashem and USHMM to have reliable references. The functionality of Wikidata was very helpful in enriching the data and make use of Linked Open Data technology. Afterwards EHRI also re-used the enriched data in her portal. Another positive aspect of Wikidata is that the platform is open for re-use by other projects, the Wikidata community and Wikipedia.



Name variants are helpful to WD users but can also be imported into the EHRI Portal, enriching EHRI taxonomies.

Adding to WD references bolsters the resource's data and directs users to reputable sources.

Following the positive experience of the Ghettos data set the same method was used for the Camps data set of EHRI. The list of Camps was expanded from 1974 to 3077 using the same sources. After importing in Wikidata they were enriched with additional descriptive information, including geographic coordinates, alternative names, associated place names, hierarchical relationships with other entries and references back to EHRI, Yad Vashem and USHMM. This enriched set was again re-used in the EHRI portal.

The collaboration with Wikidata enriched both our data and our knowledge of Linked Open Data. EHRI re-uses the enriched data sets on Ghettos and Camps and we hope that reliable and enriched information in Wikidata will also be used by other projects and Wikipedia.

Persons & VIAF

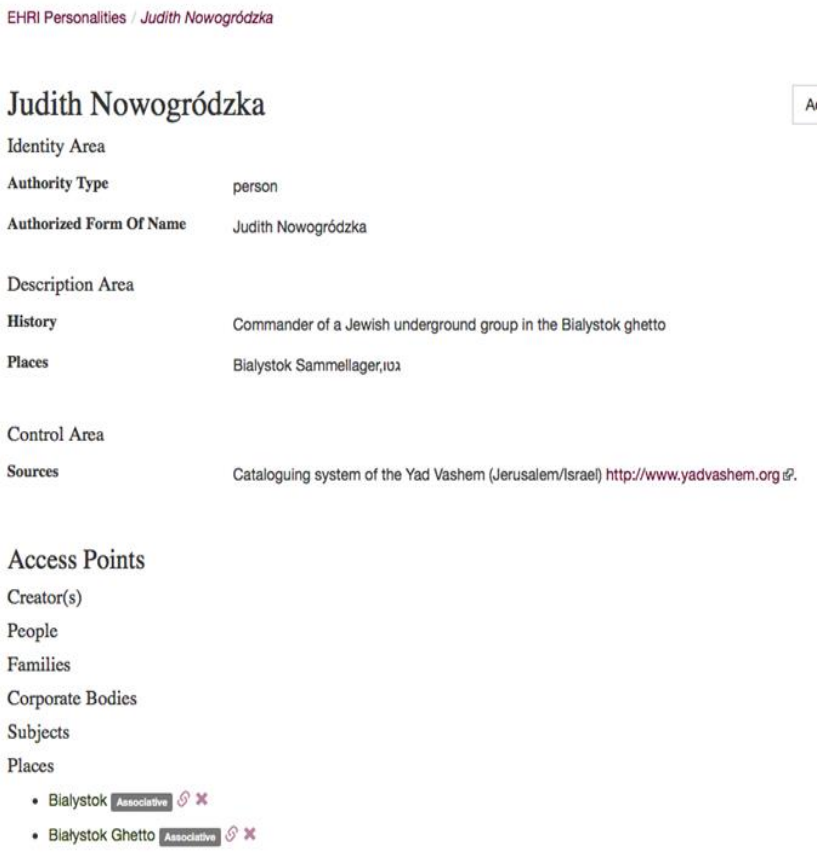
EHRI also holds authority lists on persons and corporate bodies. The persons data set focusses on the creators of the archives and collections that can be found on the EHRI portal. EHRI was searching for a reliable Internet source that holds information on persons and VIAF[4] was selected for this experiment. Though VIAF comes from the Library world and holds mostly information about authors, there were no better alternatives.

The experiment consisted of automatic matching to VIAF with persons data from Yad Vashem, CDEC and Cegesoma with manual quality check on matching results. Colleagues from these institutions made a selection of persons, manually checked them and matched this set to VIAF.

The main issues we encountered when a name is not unique or when many people carry the same name. Often there is no more additional information as birth/death dates, places or profession to disambiguate individuals and there are many spelling variants and mistakes. All of these issues made automatic matching complicated. Well known persons such as Mussolini can be easily matched automatically, but for lesser known persons or persons with a common name, the results were not good.

The result of this experiment was that from the 100 Yad Vashem, 68 were matched against entries in VIAF. There was a high ambiguity in the matching: the 68 names had 234 matches, each name was matched 3.44 times, so we proceeded with manually checking these matches. Alas the score was that 31 persons were correct and 37 were false positives. The ambiguity in cases of a correct match was sometimes higher, as one correct person was found in a set of 5/6 matches. The sets from Cegesoma and CDEC gave similar results, with the CDEC data having even much more false positives.

[4] http://viaf.org.


EHRI Personalities / *Judith Nowogródzka*

Judith Nowogródzka

Identity Area

Authority Type	person
Authorized Form Of Name	Judith Nowogródzka

Description Area

History	Commander of a Jewish underground group in the Bialystok ghetto
Places	Bialystok Sammellager,1013

Control Area

Sources	Cataloguing system of the Yad Vashem (Jerusalem/Israel) http://www.yadvashem.org.

Access Points

Creator(s)

People

Families

Corporate Bodies

Subjects

Places

- Bialystok Associative
- Białystok Ghetto Associative

As Holocaust is a sensitive topic, it proved difficult to identify the correct person when matched to (not unique) names. This experiment shows that automatic matching of persons is currently not reliable enough for the purpose of Holocaust research.

Conclusion

Lately, EHRI has set up different experiments to research if connecting the EHRI vocabularies to Linked Open Data sources on the web and using LOD to enrich and connect the vocabularies would improve the vocabularies. The results of these experiments are mixed. Working with Wikidata has proven to be useful and has enriched our knowledge, as the result of using Linked Open Data has enhanced our Camps and Ghettos lists. For enriching place names, GeoNames however has several drawbacks and is only useful when combined with a manual quality check. Finally, for persons Linked Open Data proved too infallible. When the original data is not very rich with additional information as profession or dates/places of life-events the risk of mismatches is too high and disambiguation is not always possible.