Umanistica Digitale - ISSN:2532-8816 - n.4, 2019
A. van Nispen, L. Jongma – Holocaust and World War Two Linked Open Data Developments in the Netherlands
DOI: http://doi.org/10.6092/issn.2532-8816/9048

Holocaust and World War Two Linked Open Data Developments in the Netherlands

[1]Annelies van Nispen and [2]Lizzy Jongma

[1]NIOD Institute for War-, Holocaust and Genocide Studies / European Holocaust Research Infrastructure, Herengracht 380 Amsterdam

[2]Network War Collections / Netwerk Oorlogsbronnen, Herengracht 380 Amsterdam

[1]A.vannispen@niod.knaw.nl
[2]L.Jongma@niod.knaw.nl

Abstract. NIOD, Network War Collections (Netwerk Oorlogsbronnen) and EHRI all work on connecting and making war and Holocaust collections findable and (re-)usable. And both use new technology and Linked Open Data for these goals. This paper gives an overview of the latest developments of the work done in the Netherlands. It is organized around the axis of What, Where, Who & When.

Il Network War Collection (Netwerk Oorlogsbronnen), il NIOD (Instituut voor Oorlogs-, Holocaust- en Genocidestudies) di Amsterdam, e il progetto EHRI, da diverso tempo sono impegnati nell'individuazione e collegamento di risorse documentarie sulla seconda guerra mondiale e la Shoah, per renderle ricercabili e riutilizzabili. Per il raggiungimento di questo obiettivo entrambi i progetti si avvalgono delle nuove tecnologie e dei Linked Open Data. In questo paper viene illustrato il lavoro svolto sino ad oggi.

Introduction

The NIOD, institute for War, Holocaust and Genocide Studies,[1] houses and supports the national initiative Oorlogsbronnen (War Collections) and the international research infrastructure project EHRI (European Holocaust Research Infrastructure). This paper gives an overview of where the work of NIOD, Oorlogsbronnen and, EHRI touch each other and present the latest developments on how Linked Open Data is used for connecting Holocaust and War collections in the Netherlands.

The NIOD was founded a few days after the Netherlands were liberated and its task was to

1 http://www.niod.nl/

write the History of the Netherlands under Nazi-occupation. Documentation material and archives were collected for research purposes. The NIOD still holds these archives that are witnesses of the nazi-occupation and persecution of Dutch Jews. The NIOD research department teaches the master Holocaust & Genocide studies in Amsterdam.
The international project EHRI is now currently in its second four-year project phase. EHRI is both a network of Holocaust researchers and Holocaust collections. Holocaust archives have been dispersed after the war.
EHRI[2] aims to connect all the collections it manages to discover via the EHRI portal, so that (information about) Holocaust archives are better findable for those who are researching the Holocaust. Besides the investment in connecting collections, EHRI builds on a network of Holocaust researchers with the aim to stimulate Holocaust Research.
The National program Network for War Collections (in Dutch: Oorlogsbronnen)[3] started in 2016 and aims to bring together the digitized collections of approximately 400 institutions that have holdings/collections on World War Two in the Netherlands. Oorlogsbronnen tries to connect thes collections online, making use of the latest technology. Currently, the portal site holds 45 collections and 15 collections are brought together in the People Portal.
NIOD, Oorlogsbronnen and, EHRI all work on connecting and making war and Holocaust collections findable and (re-)usable. And it uses new technology and Linked Open Data for these goals. This paper gives an overview of the latest developments. It is organized around the axis of What, Where, Who & When.

What: World War Two Thesaurus and EHRI Thesaurus

EHRI, Oorlogsbronnen and, NIOD invest in creating vocabularies that describe the knowledge domain of Holocaust and WW2. The subject terms are part of a thesaurus that enables semantic search and for Oorlogsbronnen also a knowledge graph.

World War Two Thesaurus

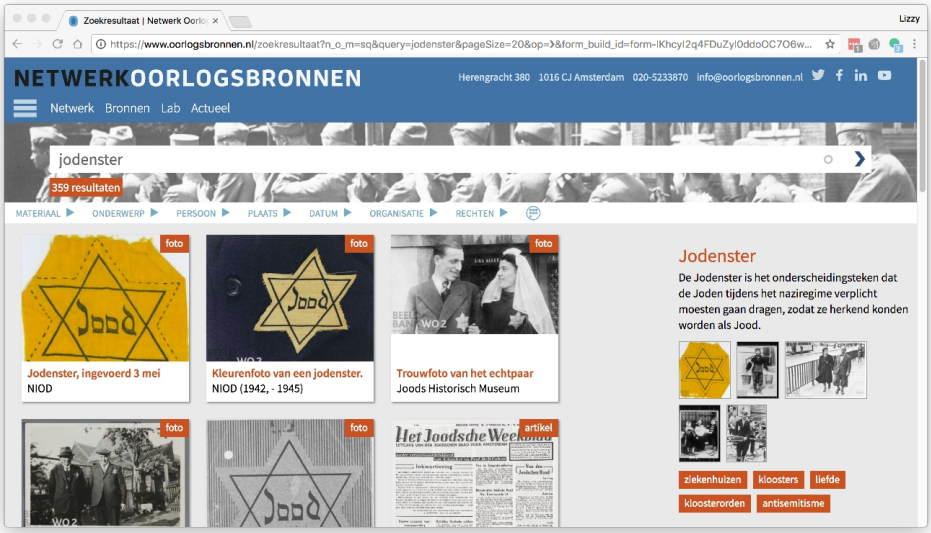
In 2016 NIOD transformed its list of subject headings into a World War Two Thesaurus. All keywords were reviewed, quality control was applied (deduplication, language check), relations were created and scope notes added. The thesaurus was migrated to a new Semantic Content Management System and made available as SKOS-XL (Simple Knowledge Organization System).[4] Elements of other ontologies were added to link to geocodes, Wikidata elements and identifiers of LOD resources. Persistent Identifiers were added to the Concepts so external partners can digitally refer to the concepts in the WW2 Thesaurus. The thesaurus is published

2 EHRI project website https://ehri-project.eu and Portal website with Holocaust collections https://portal.ehri-project.eu/
3 https://www.oorlogsbronnen.nl/
4 https://www.w3.org/2004/02/skos/

in human readable format,[5] using SPARQL and in several different LOD formats, like RDF, JSON-LD and, Turtle.[6]

Currently the WW2 Thesaurus contains over 2300 concept about politics, events, objects, places etcetera. Ranging from everyday life to the persecution of Jews (and these are also available in the EHRI Thesaurus). The thesaurus also holds a list of Camps and Ghettos in Europe and Asia. The Camps and Ghettos have been geo-referenced.

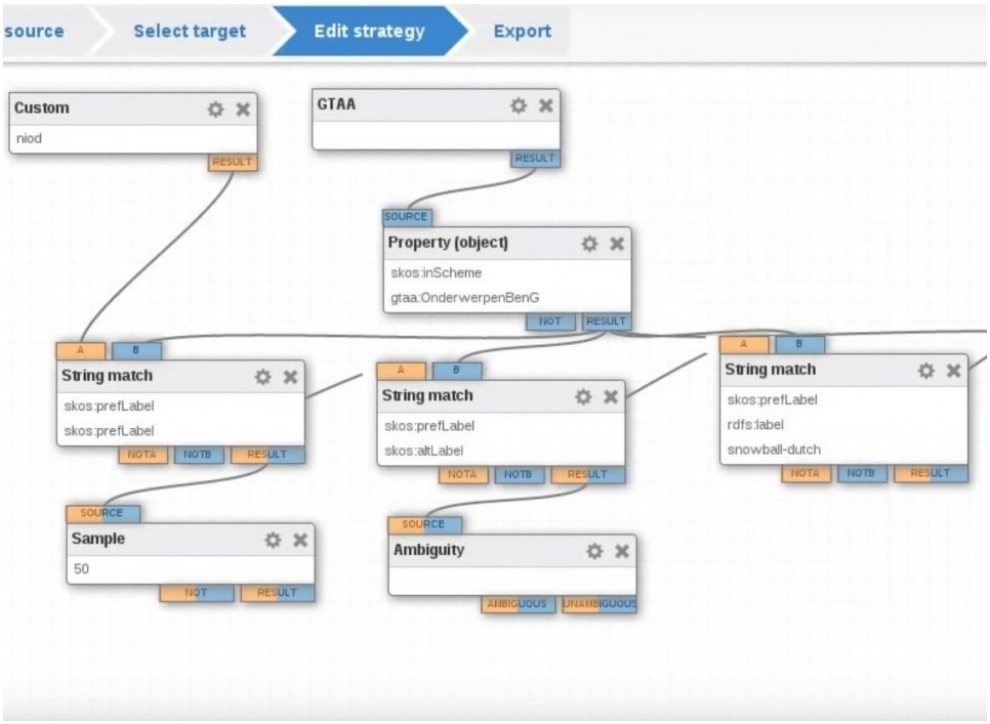
Oorlogsbronnen uses the WW2 thesaurus to match collections and make different collections of different organizations available, in context.



Picture 1: Interface of sources on website of Network War Collections. The WW2 thesaurus is also used as an info graph for additional information about specific subjects.

5 https://data.niod.nl/WO2_Thesaurus.html

6 https://data.niod.nl/WO2_Thesaurus/export/WO2_Thesaurus.ttl



Picture 2: Matching WW2 thesaurus with other Thesauri

The team of Oorlogsbronnen uses Open Refine[7] and Cultuurlink[8] to check and connect concepts in different collections with the WW2 thesaurus.

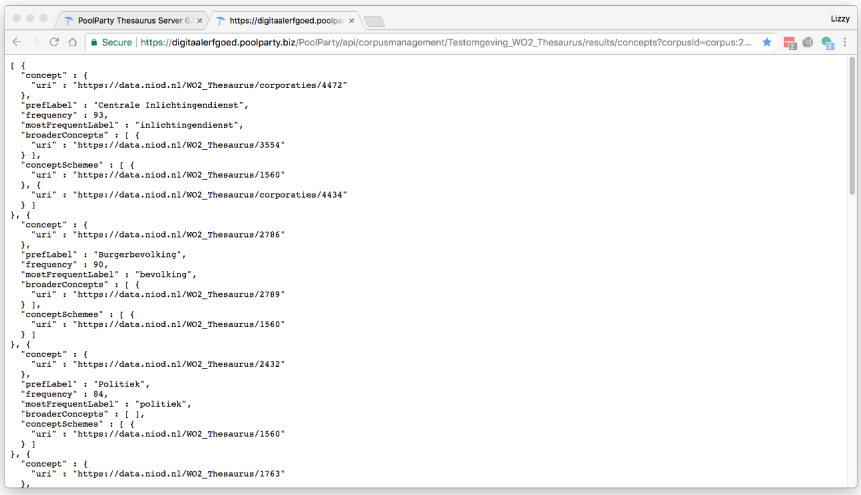
Depending on the nature of a collection (war collection or a generic collection with WW2 objects) Oorlogsbronnen was able to match up to 90% of the concepts.

The WW2 thesaurus was created in Poolparty:[9] a Linked Open Data Service built by the Semantic Web Company. Poolparty also has an API for auto-tagging and concept extraction for (full text) resources. In an auto tagging pilot project we connected OCR text and an Inventory of the National Archives in EAD to the Poolparty API and the results were promising: WW2 subjects were recognized in full-text resources and the API is able to generate reusable triples.

7 http://openrefine.org/

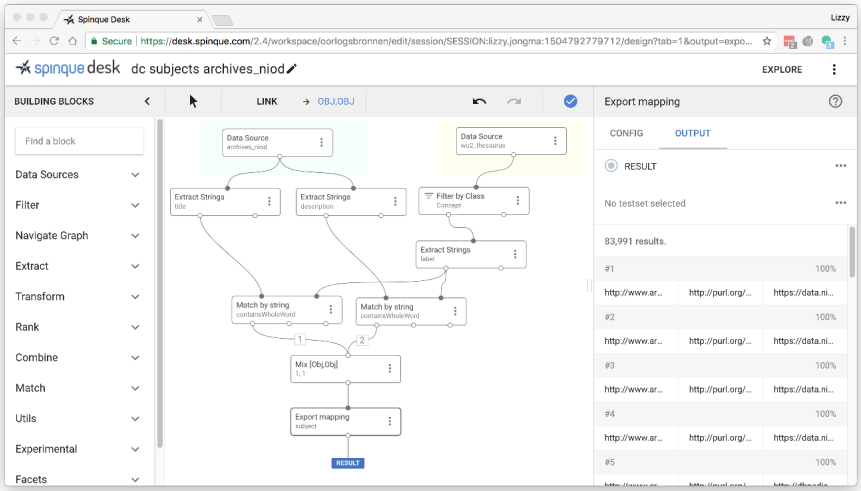
8 An online application created for the Dutch Heritage field to match Thesauri in different heritage fields. http://cultuurlink.beeldengeluid.nl/

9 https://www.poolparty.biz/



Picture 3: JSON LD result of Poolparty API

The pilot shows that we can automatically add WW2 subjects to external resources. In a second pilot we tested auto tagging within our own LOD environment (Spinque[10]). We were able to add ww2 subjects to all our war collections by matching object descriptions with the ww2 thesaurus.



Picture 4: Matching archival descriptions of the NIOD: 84.000 concepts were recognized.

10 http://www.spinque.com/

We are now able to harvest external collections in our portal and add WW2 subjects to the local descriptions. This is a big step forward in structuring a vast amount of WW2 objects and making these objects accessible for (re-)search.

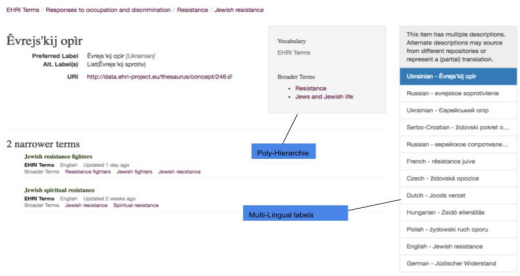
Auto-tagging helps us structure and organize full-text resources and resources without specified subjects.

EHRI Thesaurus

In the first phase of EHRI, a set of controlled vocabularies was developed to serve as a retrieval and cataloguing tool for the multilingual and highly heterogeneous data of the EHRI portal. These vocabularies were partially implemented in the first phase of the project. EHRI manages a set of controlled vocabularies, such as hierarchies of concepts and terms, person and corporate body authorities or authority lists designed for the Holocaust knowledge domain as camps, ghettos or administrative districts. 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 is to make the EHRI vocabularies efficient multilingual retrieval tools for the end users of the portal, and an efficient cataloguing and integration tool for newly ingested archival materials. EHRI wants to publish some of the vocabularies as linked open data (LOD) as well.

In the current phase of EHRI the vocabularies are in the process of quality improvement (deduplication, merging, adding multilingual labels, consistency checks, multiple parent relations, etc.) and increase their coverage. To continue the work, improve and enrich the existing terms, add new terms, disambiguate and remove the mistakes EHRI needed a tool designed and implemented for vocabulary management activities. it was decided that the VocBench tool[11] will be used as a thesaurus management system. This synchronizes weekly with the EHRI portal. In the EHRI portal the subject terms are currently not available for the public.



Picture 5: EHRI terms in EHRI portal

11 http://vocbench.uniroma2.it/

In EHRI-1 a Persecution and Holocaust subset of NIOD's subject term list was created and sent to EHRI. This was partially included in the portal. In EHRI-2 this list has been analyzed and matched with the EHRI thesaurus. Where possible, terms were merged with the EHRI thesaurus. In the future our aim is to connect the relevant WW2-thesaurus terms to the EHRI Thesaurus through LOD. The EHRI thesaurus has multi-lingual labels and sometimes additional information that could enrich the Holocaust related terms in the World War Two Thesaurus.

There are currently no plans to translate the World War Two Thesaurus in other languages.

Where: EHRI Ghettos and Camps

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 Ghettos and Camps & Traces of War in the Netherlands

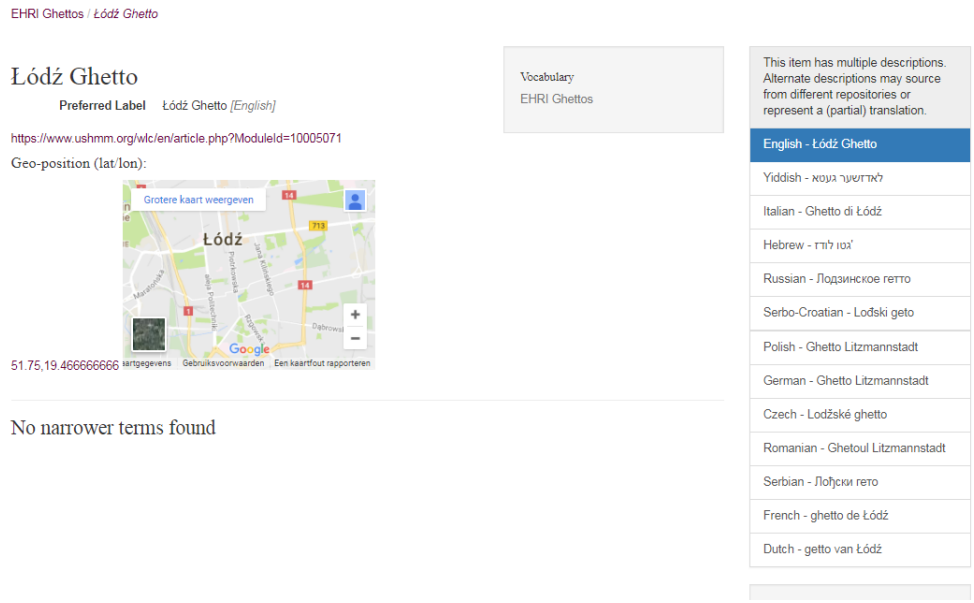
EHRI has several authority files for places. It works on an authoritative overview of Ghettos and Camps. These vocabularies are mostly based on decennial long work that Yad Vashem and USHMM have done in their encyclopedias[12] of Camps and Ghettos. And it has a places vocabulary that is linked to Geonames. These vocabularies are being used for cataloguing and retrieval purposes.

In spring 2017 EHRI started a pilot with Ghettos and Wikidata. The aim was to experiment with working with Wikidata and see if we could create a win-win for both the EHRI project and Wikidata/Wiki-community. Our expectation was that EHRI could help on creating a reliable overview of Ghettos for Wikidata that can be used more widely in the Wiki and LOD-communities. The incentive for EHRI was that Wikidata has better developed functionality to enrich descriptions and make use of Linked Open Data vocabularies. The enriched Ghetto-information is being used by the EHRI portal as well.

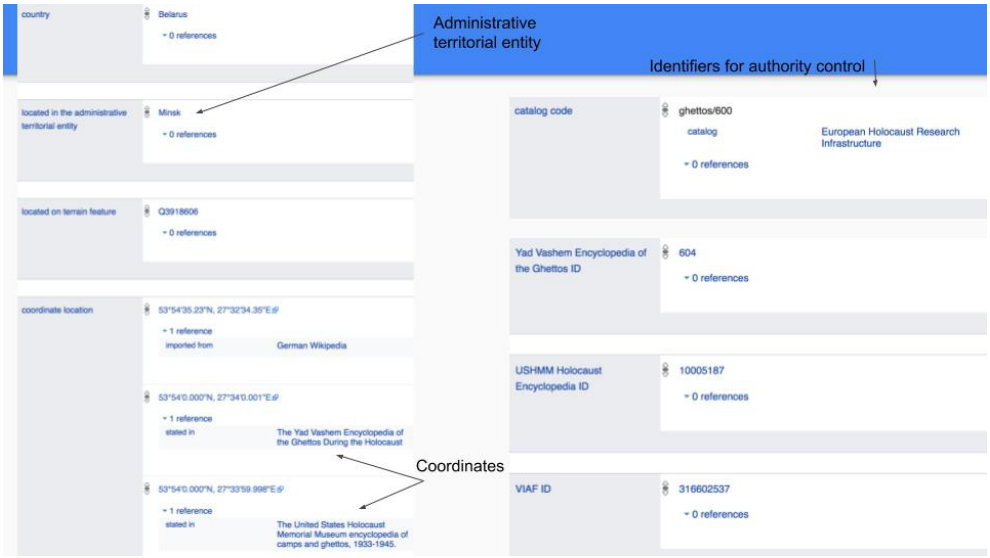
At the start of the pilot there was reliable information about 85 ghettos available in Wikidata. EHRI has imported its Ghettos vocabulary into Wikidata and started to enrich the descriptions with multi-lingual labels, geographic information (including Administrative Districts, a separate EHRI Authority list) and linking it to reliable sources from USHMM and Yad Vashem and of course EHRI. In June 2017 1368 Ghettos in Wikidata are available for (re-)use.
In the upcoming period we hope that the wiki & LOD-community will make use of the

12 Encyclopedia of Camps and Ghettos, 1933-1945 https://www.ushmm.org/research/publications/encyclopedia-camps-ghettos; example from Yad Vashem Lodz Ghetto http://www.yadvashem.org/yv/he/research/ghettos_encyclopedia/ghetto_details.asp?cid=516

Ghettos. The EHRI team started work on improving and enriching the Camps vocabulary.



Picture 6: Lodz Ghetto in EHRI



Picture 7: Lodz Ghetto in Wikidata

Traces of War in the Netherlands

In 2016 Oorlogsbronnen investigated options to enrich and improve the searchability of collections by adding geocodes to names of places mentioned in the metadata. The metadata format that is was used to create interoperability was Dublin Core. Currently Oorlogsbronnen working on integrating richer metadata, as EAD.

The project team investigated several options to detect places in present Dublin Core metadata:

- Places mentioned in the Dublin Core field dc:coverage and dcterms:spatial
- Places mentioned in the Dublin Core field dc:subject
- Places mentioned in the descriptive metadata dc:title and dc:description.

The outcomes of this project are that:

- In the Netherlands Heritage Institutes collaborate in building a thesaurus of Historical buildings and streets, called the BAG. This specialized thesaurus works best, but is -unfortunately- limited in size. In many cases GeoNames is the best and easiest to use thesaurus for places.
- A lot of historical data is missing in geographical data: you can best help yourself and help others by adding historical names of places and historical places. In this project The Dutch Indies, the Soviet Union, Yugoslavia, and prison Oranje Hotel were added.
- Names of places in dc:coverage are the easiest to match and create the least amount of false positives (when a computer thinks that names match, but humans know that the names don't match). But, still 25-30% cannot be matched fully automatically.
- Adding geo data to objects after aggregation helps to demonstrate new ways of searching and presenting objects, but adding relevant information outside the Heritage Institution creates new problems: the additional information is separated.
- From the original metadata and limited in usability and reusability. Geodata should be added to the original metadata by Heritage Institutions. A report with analysis and methodology is available on the Oorlogsbronnen Github (only in Dutch).[13]

In the summer of 2017 Oorlogsbronnen will build a pipeline to add geolocations to all objects in the portal (with named places in dc:coverage and dcterms:spatial) and we will test sharing geodata with our network partners.

13 Data and reports are available at the Oorlogsbronnen github: https://github.com/NetwerkOorlogsBronnen/pilot-geocoderen

Who: People of World War Two

Most people searching for information about the Second World War are searching for names: names of victims, collaborators or perpetrators. And usually people are looking for relatives. Searches for information on persons are a large part of search requests. Both EHRI and Oorlogsbronnen want to facilitate these searches by structuring biographical data (using LOD) and making it available to the public.

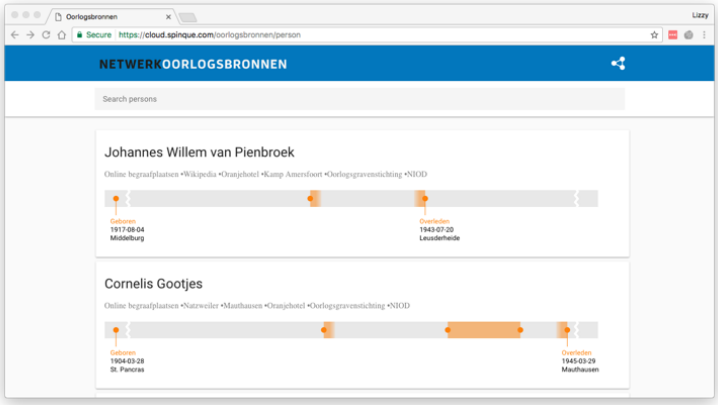
People Portal

Oorlogsbronnen is currently working on People of WW2/the People Portal. As a stand-alone portal site about people involved in the Second World War in the Netherlands and as part of the collections portal.

Goal of this project is to create a LOD set with prosopographical and (minimal) biographical data of people affected. The set will consist of statesmen and politicians, military men, resistance fighters, collaborators, victims of persecution, forced laborers and civilians.

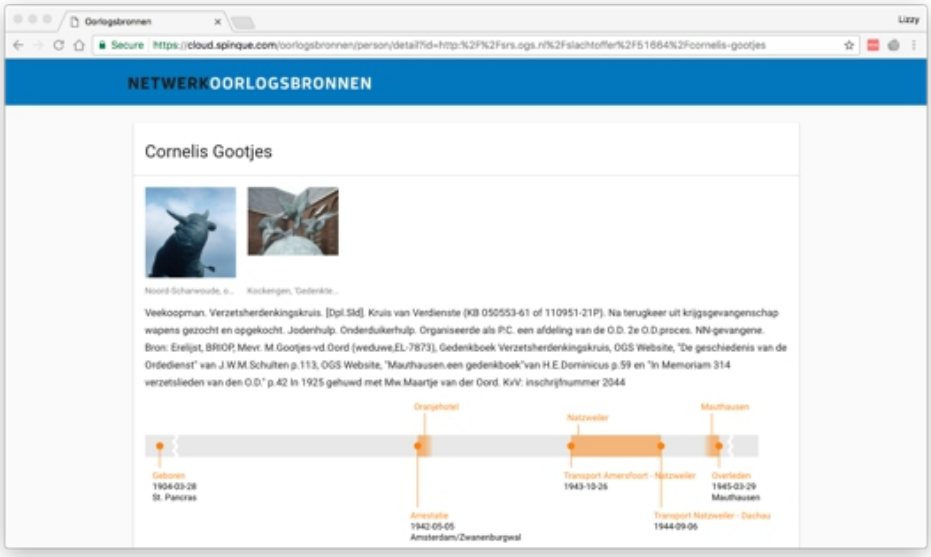
So far Oorlogsbronnen collected approximately 600.000 names from 10 different resources, but the dataset will probably grow to more than a million names. There is a big difference between the mentioning of a name and an individual person: names (entities detected by humans as a name) can be misspelled, parts of names can be left out or abbreviated, names change over countries, languages and time. One name can point to several different individuals. To detect a Person we need data about date and place of birth, date and place of death etc. And we need different resources to compose the lives of individuals.

But, even though most digitized resources about WW2 are limited, there is good news: if we combine several resources with limited data then we can start detecting different Persons.

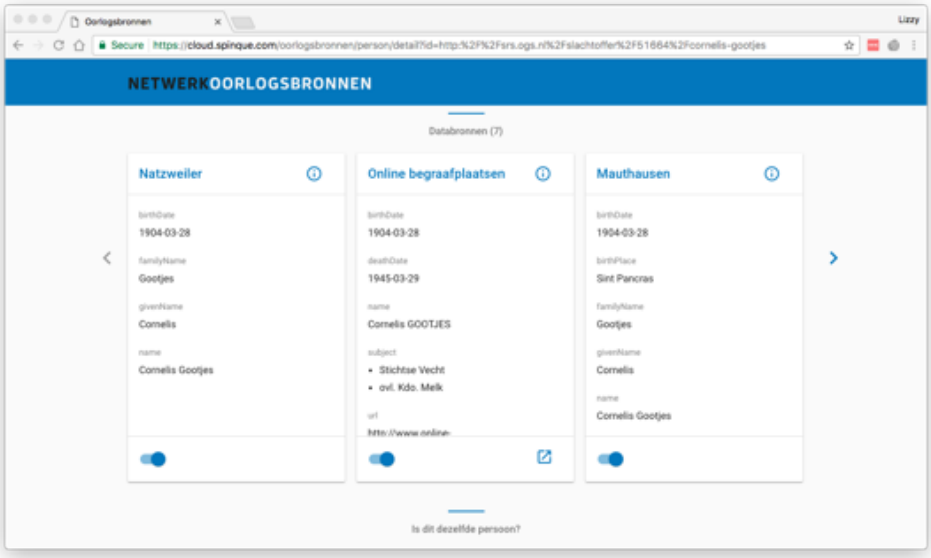


Picture 8: Screenshot of the Demo version of the People Portal

In the Oorlogsbronnen Demonstrator we already found approximately 100.000 individuals in 10 resources. We can also create a timeline for these Persons and place their lives/events on a map.



Picture 9: Screenshot of the Demo version of the People Portal



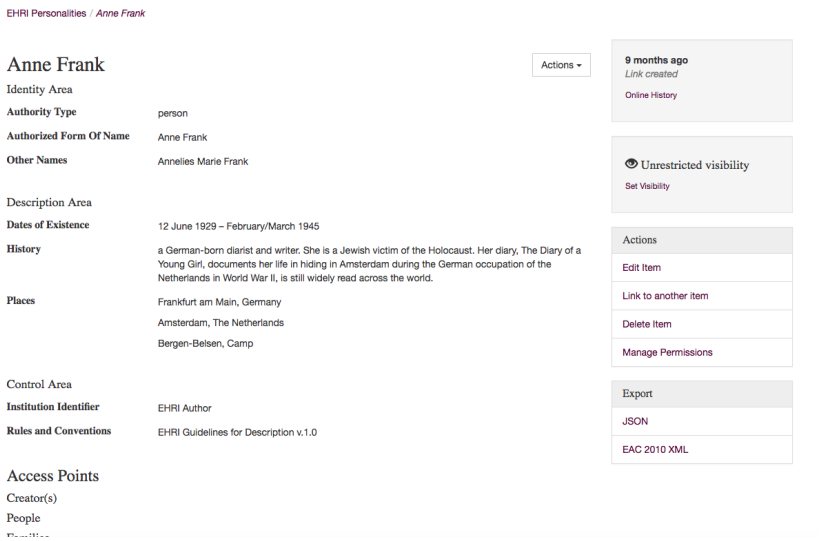
Picture 10: Screenshot of the Demo version of the People Portal

A big issue, when working with biographical data, is privacy. We create new digital data about people and write new algorithms in order to learn new things about people and about human

behavior.
But, do we want all of our data being used for all kinds of data analyses? Do we want to share our entire life with everyone?
People are becoming more concerned about their own privacy and the privacy of relatives. And privacy laws are becoming stricter. A lot of Heritage Institutions have limited knowledge about constantly changing privacy laws and don't want to get into legal problems.
So, even though the institutions digitize more and more: less and less becomes available online because of Copyright or Privacy restrictions.
Oorlogsbronnen asked Kennisland, a Dutch Think Tank on Social Innovation and Smart Societies to explain and document Privacy Laws for Heritage Institutions[14]. Conclusions are that privacy Laws apply to living people. But, if data about a dead person violates the privacy of a living person then this data is restricted too.
Institutions are allowed to process personal data, but need to ask permission of the living (or good arguments why it is impossible to get permission). The data needs to be processed and stored on well protected machines and calamity procedures implemented.
So, prosopographical research into the lives of people affected by WW2 is possible. But, due to privacy issues we cannot share (all) data with external partners.

EHRIs Historical Agents

EHRI is also working on Person information. This will be one of the last vocabularies that will be improved. There are currently around 12.000 persons described in the EHRI Personalities. The Terezin victims list, personalities coming from Yad Vashem form the main body. From the NIOD keywords persons were extracted and added to the EHRI personalities, *e.g.* Anne Frank.



Picture 11: Anne Frank in EHRI personalities

14 https://www.oorlogsbronnen.nl/nieuws/verwerken-en-publiceren-van-persoonsgegevens-uit-40-45-onderzocht

In 2018 work will start to add, improve and enrich personalities in EHRI. The strongest focus of the EHRI personalities is to describe the creators of the archives. The EHRI personalities already include a subsection of Dutch personalities. We foresee that the work that Oorlogsbronnen is doing on Dutch Persons can also strengthen EHRI. EHRI could reuse the Holocaust sources or enriched persons corpus that Oorlogsbronnen is creating.

Where: Events

Authors are both members of the Events Working Group http://www.ehumanities.nl/events-working-group/. The work of this group is motivated by the 1) demands for facilitating a deeper understanding of online cultural heritage collections, and by the fact that 2) events emerged as a key element in the representation of data in areas such as history, cultural heritage, and multimedia. We bring together computer scientists, computational linguists and humanities and social sciences scholars in order to *build upon and expand the results in existing research communities*, *e.g.* NLP, Information Retrieval, Semantic Web, Social Web Analytics, Multimedia analysis, and *provide structure and deeper understanding* in history, media, journalism and cultural heritage research, with a specific focus on how events are used as a key concept for representing knowledge and organizing media in online web collections. The ultimate goal is to distill a **research and application roadmaps** for events in Cultural Heritage, e.g. achieving a social consensus on processes, identify practical standards and protocols, defining the infrastructure needed.

NOB is a use case for this WG and in future EHRI can serve as one as well.

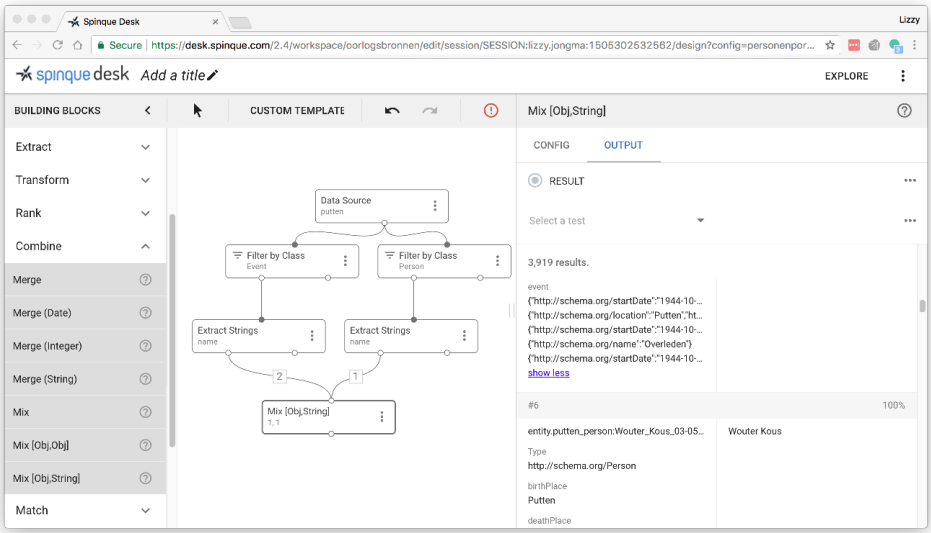
A case where this is present is the Bunkerdrama of Camp Vught (January 15, 1944): 74 women were crammed into a bunker for 24 hours as punishment. 10 women died.

We try to create semantic models in which, from macro-level perspective, we can connect WW2 to the Bunkerdrama and to the death of 10 women. And from a micro level perspective, we want to connect the arrest, detention and death of a person to the Bunkerdrama.

Our goal is to create LOD models to describe and structure events of different levels and then to be able to detect them in resources.

In the Linked Open Data communities a lot of research and development was put into ontologies for thesauri, geo data etc. But little has been done in the field of (historical) event modeling.

The Event Working Group is testing the Simple Event Model (SEM) in close collaboration with the developers of the model. Oorlogsbronnen is also testing schema.org for event modeling. Even though schema.org is less structured, it is flexible, well used and well documented. And hopefully will help improve search engine results for information in the portal of Oorlogsbronnen.



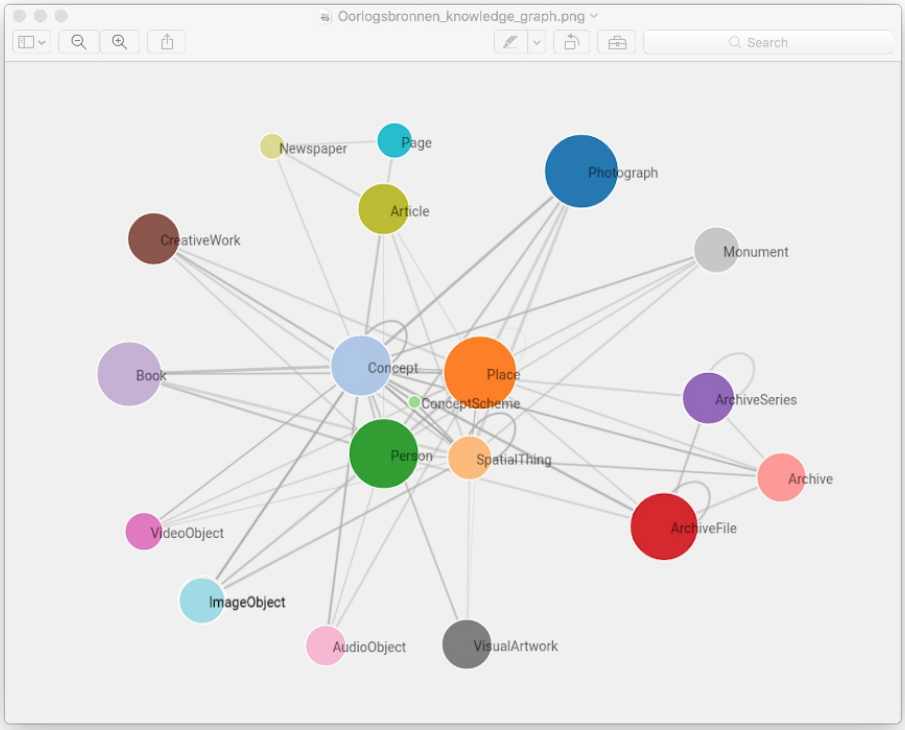
Picture 12: Sample result of Event Modelling with schema.org. Data Source is the website http://www.oktober44.nl/ about the Razzia of Putten

Conclusion and future work

Only 7% of all relevant information in the Netherlands is digitized: a lot of data is becoming available and digitization of heritage is speeding up but a lot of information is not available and reusable yet.

Copyright laws and privacy protection laws are not helping Heritage Institutions for 20th Century material, which all Holocaust archives are. They restrict us in our possibilities to use the full potential of the digital world and the internet. A long term effect could be that recent parts of our History are hidden away and forgotten, because of legal issues.

Heritage institutions and aggregators are moving away from the old “copy-collections” model. Using Linked Open Data techniques and creating references to thesauri helps to connect and explain data. Single, meaningless items are remodeled into interlinked units sources can become more (re)connected, new and old patterns may become visible.



Picture 13: Visualisation of RDF elements in de Oorlogsbronnen portal shows the shift from copy-collection to who, what, where, when driven connections.

Oorlogsbronnen and EHRI are connecting resources on four basic historical questions: what, where, who, what happened? Computers and big data can do a lot to structure and improve information. To do new research, ask new questions. LOD technology can improve searching, finding (and hopefully understanding) resources greatly. It enables new ways of interconnecting sources that are poorly annotated, unstructured or described with local terminologies. Structuring and combining resources can help users and empower research.
But, human interventions are still necessary: computers can draw a lot of wrong conclusions. Aggregators currently work on improving data but the improved data remains separated from the original data. Heritage institutions need to become involved in the technical developments.

Acknowledgments

We like to thank the organizers and participants of the EHRI Workshop Data Sharing, Holocaust Documentation, Digital Humanities: Best Practices, Case Studies, Benefits.