



**European Holocaust Research Infrastructure  
H2020-INFRAIA-2014-2015  
GA no. 654164**

**D.11.2**

**Road Map Domain Vocabularies**

**Kepa J. Rodriguez (YV)  
Vladimir Alexiev (ONTO)  
Laura Brazzo (CDEC)  
Charles Riondet (INRIA)  
Yael Gherman (YV)  
Reto Speck (NIOD)**

**Start: [1]**

**Due: [10]**

**Actual: [13]**



**EHRI is funded by the European Union**

## Document Information

Project URL	<a href="http://www.ehri-project.eu">www.ehri-project.eu</a>
Document URL	
Deliverable	11.2 Road Map Domain Vocabularies
Work Package	11
Lead Beneficiary	6 / YV
Relevant Milestones	MS1
Dissemination level	PU
Contact Person	Kepa J. Rodriguez kepa.rodriguez@yadvashem.org.il +972 644 3402
Abstract (for dissemination)	<p>In this document we present an analysis of the current implementation of the vocabularies EHRI developed in its first phase as well as a strategy to improve them and increase their coverage. 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. Although it has a lower priority for the project, we further project to expose some of the vocabularies as linked open data (LOD).</p> <p>An important part of the curation of the EHRI vocabularies is the validation of modifications by content specialists. We will therefore establish an editorial board which will be charged to validate new concepts and authorities and oversee changes in the vocabularies.</p>
Management Summary	<p>In the first phase of the project, EHRI developed a set of controlled vocabularies with the aim of improving retrieval of the multilingual and highly heterogeneous data of the portal. These vocabularies were partially implemented in the first phase of the project.</p> <p>In this document we present a detailed analysis of the implementation realised in EHRI-1 and the results that were thus achieved. Since the main purpose of the vocabularies was the retrieval of descriptions of archival units, one of the most important aspects we evaluated is the current use of the vocabularies to add access points to such descriptions imported into the portal.</p> <p>We further present a plan to enhance the coverage of the vocabularies by increasing the number of concepts and authorities, increasing the linkage between the different vocabulary sets, and linking access points provided by the Collection Holding Institutions (CHIs) to the EHRI vocabularies.</p> <p>The strategy we intend to adopt is the opposite of the strategy</p>

employed in EHRI-1. In the previous phase of the project, the vocabularies were developed before the data were imported into the portal. The starting point of the vocabularies was the knowledge of the Holocaust specialists of the EHRI partners. In the second phase of the project, we will implement a data-driven strategy, learning and extracting knowledge from the data and using it to extend the coverage of the different vocabularies and to find relationships between them. We will use this information to infer links between the descriptions of units and the vocabularies.

Since the modifications have to be validated by content specialists, our approach includes the integration of an editorial board in the workflow. The board will validate modifications realised in the vocabularies, thereby ensuring the quality and consistency of the results.

Although the exposition of the vocabularies for external consumers has a lower priority for the project, we will attempt to expose a part of the thesaurus as Linked Open Data (LOD). The extent to which we will publishing LOD will depend on the remaining resources after the other goals have been achieved.

Finally this deliverable presents a fine-grained break-down of the necessary tasks, an estimation of the required resources, the partners that will be responsible for the various tasks, and a timetable for their execution.

## Table of Contents

Document Information .....	2
1 Introduction .....	6
1.1 Description of the document.....	7
2 State of the art: controlled vocabularies in other projects.....	8
2.1 Interlinking approach.....	9
2.1.1 PELAGIOS.....	9
2.1.2 CDEC Digital Library .....	10
2.1.3 LIPARM Project.....	11
2.2 Aggregation Approach.....	12
2.2.1 ArchivesHub.....	12
2.2.2 Israel Archives Network.....	13
2.2.3 CENDARI.....	13
2.2.4 Getty Vocabularies.....	14
2.3 Conclusions.....	16
2.3.1 Comparison of the two approaches and relationship to EHRI.....	18
3 Existing EHRI controlled vocabularies .....	20
3.1 Vocabularies imported into the portal .....	20
3.1.1 EHRI-1 vocabularies .....	20
3.1.2 Research guides vocabularies .....	22
3.1.3 Additional vocabularies .....	23
3.1.4 The EHRI Thesaurus as a retrieval tool.....	23
3.1.5 EHRI Thesaurus Quality.....	30
3.2 Vocabularies published as Linked Open Data .....	32
3.2.1 Problems detected in the LOD set.....	33
4 Structure of the EHRI vocabulary "universe" .....	36
4.1 Centralized approach .....	36
4.2 Federated approach.....	37
4.3 Combinatorial approach .....	38
4.4 Conclusions.....	40
5 Dataflows .....	41
6 Thesaurus Editorial Policies .....	42
6.1 EHRI Thesaurus Editorial Board and Workflows .....	42
7 Transformation and curation of vocabularies .....	44
7.1 Transformation and Curation of existing vocabularies .....	44
7.1.1 Concepts (Subjects).....	44
7.1.2 Places .....	45
7.1.3 Camps, Ghettos, Murder Sites .....	48
7.1.4 Person entities .....	50
7.1.5 Corporate bodies.....	52
7.1.6 Events .....	52
7.1.7 Administrative districts.....	53
7.2 Transformation and Curation of new introduced entries/instances.....	53
7.2.1 Concepts.....	53

---

7.2.2	Person entities .....	53
7.2.3	Corporate bodies.....	53
2.1.1	Events .....	54
7.3	Creation and modelling of new vocabularies .....	54
8	Publishing the Authorities .....	55
8.1	EAC CPF Formats.....	55
8.2	Linked Open Data .....	55
8.3	Auto-Completion .....	56
9	Tools and Applications .....	57
9.1	Edition on the administrative interface of the portal .....	57
9.2	Wikidata as a Semantic Integration and Editing Platform .....	57
9.2.1	Wikidata for Editing .....	57
9.2.2	Wikidata for Integration and Coreferencing .....	58
9.2.3	Using Wikidata for Camps and Ghettos.....	59
9.3	VocBench.....	60
9.3.1	VocBench Users .....	60
9.3.2	VocBench Features.....	61
9.4	SKOS Visualization (SKOSPlay) .....	62
10	Effort Estimation and Time Table .....	63
	Appendix .....	67
	Glossary .....	74
	References.....	75

# 1 Introduction

The EHRI portal contains very heterogeneous data. The data is not only highly multilingual, but it has been catalogued in very different ways, using different conventions to assign keywords or access points to the collections. That makes the implementation of good information retrieval tools a challenging task. For instance, free text search has difficulties with different spellings<sup>1</sup>, and difficulties to handle synonymy and multilingual data (different labels for the same concept).

In the first phase of the project, EHRI developed a thesaurus or set of controlled vocabularies (concepts and authorities) with the aim of addressing the retrieval problem ([Gertner et al 2015]), and partially implemented it. We can cluster the retrieval related function of the EHRI thesaurus into three groups:

1. To serve as the main tool for multilingual information retrieval. Description units can be retrieved with a single query regardless of the language in which they are written, if they are linked to the same item in the vocabulary.
2. To serve as a cataloguing and integration tool for newly ingested or manually introduced description units. Its use in the cataloguing of new descriptions in the portal will interlink the description units with other units in the portal, making them retrievable.
3. To serve as knowledge basis for domain specific development and training of NLP models and tools. Standard NLP tools have been trained and tested using standard corpora, most of them consisting of annotated newspaper articles or pages of Wikipedia. Although they have a good performance in the extraction of names of people and actually existing cities, they have problems detecting historical entities.

Not all the possible involved subtasks have the same priority for the project. The Project Management Board selected and ranked by priority the following subtasks:

- Improve the functionality of the vocabularies in order for it to serve as a retrieval tool
- Align and map of Collection Holding Institutions' (CHIs) Controlled Vocabularies/Thesauri to the EHRI Controlled Vocabularies/Thesaurus
- Develop and implement vocabulary management tools and workflows
- Include missing vocabularies and improve the functionality of the vocabularies in order to improve their values as cataloguing tools

---

<sup>1</sup> Different spellings of names are quite common in the EHRI portal. One of the reasons is that often entities are transliterated into Latin alphabet from Hebrew/Yiddish and Cyrillic. Another factor that increases this problem is the use of language specific diacritics. Diacritics are not a problem for free text search, if they are not transliterated, but if the writer tries to interpret and transliterate them, it becomes a source of multiple alternative spellings of the same word.

The following subtasks were ranked with a lower priority for our work package, but are relevant for the development of other work packages such as WP10 (Resource Identification and Integration Workflows) and WP13 (Research data infrastructures for Holocaust material):

- Create a common workflow for publishing Linked Open Data from the EHRI vocabularies, making EHRI Controlled Vocabularies available for third parties
- Enhance the CHI-provided descriptions by automated entity extraction of existing data before integration

In this document we present an analysis of the current implementation of the vocabularies and strategies to improve the current sets and increase their coverage. The goal is to make the EHRI domain vocabularies an efficient retrieval tool for our users.

## 1.1 Description of the document

Section 2 presents how other projects manage and use controlled vocabularies for the ingestion and presentation of data. We classify the projects into two groups: interlinking, where vocabularies used in different data sources are linked; and aggregation, where a central vocabulary or set of vocabularies is enriched with entries provided by the data providers or extracted from the data.

Section 3 describes the current EHRI thesaurus and its usage in the portal, after which it points out critical issues which affect its use as a retrieval tool.

Section 4 discusses three approaches to organizing controlled vocabularies: centralized, federated and a combination of both, the combinatorial approach.

Section 5 presents the data flows to handle access points of new data during the ingestion process. Section 6 outlines the importance of creating an editorial board to manage modifications in the content of the vocabularies.

Section 7 presents and discusses curation strategies for the existing EHRI vocabularies and gives examples of new vocabularies, which could be created if requested by the project. Section 8 describes the proposed formats for the publication of authority files.

Section 9 describes the functionality of three tools: the administrative interface for Events and EAC<sup>2</sup> files; VocBench for thesaurus concepts; and WikiData for ghettos and concentration camps.

Finally, section 10 shows a division of person months between the partners for the necessary activities needed to fulfil the recommendations of this document, as well as a provisional timeline outlining the sequence of these activities.

---

<sup>2</sup> Encoded Archival Context – Corporate bodies, Persons, Families

## 2 State of the art: controlled vocabularies in other projects

The EHRI project is engaged, since its first phase, in the construction and implementation of controlled vocabularies. This task is at the heart of the project because it is chiefly through controlled vocabularies that effective and precise retrieval of information is possible.

These vocabularies will be published by EHRI, possibly in Linked Data format, in order to facilitate the (semantic) interoperability with other information systems and to foster sharing and re-use of data.

In the framework of the (digital) humanities, the wide proliferation of data has led to the growing need of controlled vocabularies and authoritative lists, in order to both provide web users with retrieval tools that make their searches as precise as possible, and at the same time, to overcome problems related to the use of different terminologies.

In fact, concerning subjects or entities such as people or places, the LCSH<sup>3</sup>, VIAF<sup>4</sup>, Geonames<sup>5</sup> datasets are able to greatly meet the needs. However, in cases of specific domain areas of research, the need to go deeper into the descriptions often leads to the creation of specific domain vocabularies. The EHRI project, like other similar projects, has been involved in the creation of such domain vocabularies.

It should be noted here that the development of controlled vocabularies and thesauri is strictly connected to activities relating to data integration. Data integration strategies have a marked influence on the creation (as well as adoption) of controlled vocabularies.

A brief analysis of data integration strategies and approaches adopted by some other Research Infrastructure (RI) projects is provided here, and will serve to contextualise our own work in the wider RI landscape.

We have considered seven RI projects: Pelagios, CDEC Digital Library, LIPARM, ArchivesHub, Israel Archives Network (IAN), Cendari, and the Getty Thesaurus. All of them have points of connection or similarities with the EHRI project, whether they be general aims (integration of archives and archival descriptions), topics or issues to be overcome.

Among these seven projects, we have identified two main tendencies or better approaches to integration: the interlinking approach and the aggregation approach.

The interlinking approach finds its core in linking each other's already existing and autonomous datasets through specific and common entities; the aggregation approach employs the extraction and import of data from a variety of repositories to a unique architecture and, in some cases, the re-modelling of the ingested data on the basis of the project's data-model. The projects Pelagios, LIPARM and CDEC Digital Library provide examples of the interlinking approach; Archiveshub, IAN, Ariadne, Cendari and Getty Thesaurus are examples of the aggregation approach.

---

<sup>3</sup> <http://id.loc.gov/authorities/subjects.html>

<sup>4</sup> VIAF "virtually combines multiple LAM (Library Archives Museum) name authority files into a single name authority service", see <http://viaf.org/viaf/data/>

<sup>5</sup> <http://www.geonames.org/about.html>



## 2.1 Interlinking approach

### 2.1.1 PELAGIOS

[PELAGIOS](#) (Pelagios: Enable Linked Ancient Geodata In Open Systems) is a community network that facilitates the mapping and linking of online resources on the Greco-Roman period, exploiting LOD technologies.

The entities of type “Place” are the core of the project, and the [Pleiades dataset](#)<sup>6</sup> is used for the making of the map.

The Pleiades data structure is based on three conceptual entities: [Place, Location, and Name](#). The [Pleiades data model](#) shows that the Pleiades content is organized in six defined classes: Place Resources, Name Resources, Location Resources, Reference Citations, Temporal Attestation, and Positional Accuracy Assessments.

Eight [controlled vocabularies](#) have been employed by Pleiades to define:

1. [Association certainty](#)
2. [Attestation confidence](#)
3. [Language and script](#)
4. [Name accuracy](#)
5. [Name completeness](#)
6. [Name Types](#)
7. [Feature \(or Place\) categories](#)
8. [Time Periods](#)

Ancient places from Pleiades have been georeferenced using georeferencing tools; more accurate identifications and labelling of locations has been entrusted to individuals and online communities.

From the [PELAGIOS map](#) the user is redirected to the list of the associated resources ([references](#)) as well as to the URI of the sought place (for example <http://pleiades.stoa.org/places/42307>).

To search resources associated to Places, PELAGIOS developed a search engine, [Peripleo](#), which provides machine access to data, starting from a model based on three types of entities: Items (archaeological artefacts, text, photographs), Places (related to the items), and Datasets (collections of items).

In connection with the PELAGIOS project, it is worth mentioning the ongoing project [SNAP DRGN](#) (Standard for Networking Ancient Prosopographies: Data and Relations in Greco-Roman Names) aimed at building a virtual authority list for ancient persons. Person data come from three datasets: the Lexicon of Greek Personal Names (persons mentioned in ancient Greek texts); Trismegistos (names and persons from Egyptian papyri); and Prosopographia Imperii Romani (senators and other elites from the first three centuries of the Roman Empire).

---

<sup>6</sup> Digitalization of the Barrington Atlas

## 2.1.2 CDEC Digital Library

The [CDEC Digital Library](#) is an Italian project aimed at the integration of data produced by the working areas of the CDEC Foundation. The integration strategy has been to connect resources while preserving the autonomy of the original databases and repositories. For this task LOD technologies have been adopted.

The integration process has its focal point in the entity “Person” with which all the other entities are interlinked (and through which they can be accessed).

Controlled vocabularies come from the classes of the [domain ontology](#) developed to describe the persecution experience of the Jews in Italy during the Holocaust. In addition to the controlled vocabulary about persons, there are controlled vocabularies about:

- places (arrest places, detention places, gathering places)
- Nazi camps (sub-camps included)
- massacres
- prisons
- convoys

All vocabularies are interlinked.

Data about persons and persecution come from the CDEC Database of Italian Shoah Victims' Names.

The corporate bodies vocabulary has been created using data from EAC-CPF. New items ingested in the current controlled vocabularies are added manually. Items from persons, corporate bodies and places vocabularies are linked with the description of the CDEC archival collections through automatic reasoning.

Using the “same as” function, persons, places and Nazi camps IRIs are interlinked (places and Nazi camps automatically, persons manually) with VIAF, Geonames, DBpedia. Persons IRIs are also interlinked (when possible) with those from the Italian Chamber of Deputies dataset ([dati.camera.it](http://dati.camera.it)), as well as the Italian Central State Archive ([dati.acs.it](http://dati.acs.it)) dataset.

Vocabulary items are access points to resources published in the CDEC Digital Library. An advanced search engine is available to search for persons starting from data about persecution:

- Arrest place
- Prison
- Gathering place
- Deportation Nazi camp
- Fate
- Death type
- Massacre

Person types: Victim, Author (library author), Photography author, Index item, and Donor are access points on the portal.

Resources linked to external datasets (at the moment only those associated with persons) are currently displayed in the CDEC Digital Library (for example <http://digital-library.cdec.it/cdec-web/person/detail/pcv'erson-5320/morpurgo-elio.html> ).

### 2.1.3 LIPARM Project

The main aim of the [LIPARM \(Linking Parliamentary Records through Metadata\)](#) project<sup>7</sup> is integrating parliamentary metadata. The idea behind this project is the creation of a unique interface and point of access for the digitalized UK parliamentary proceedings that were scattered throughout several points of access and platforms.

A specific metadata schema, the Parliamentary Metadata Language Schema ([PML schema](#)), has been created for supplementing and linking together already existing schemas, providing features such as list of names, biographic information, biographies, and subjects.

XML identifiers have been used to link the seven top-level components of the PLM schema (Unit, Functions, Persons, Calendar objects, Proceeding objects; Proceeding groups, Vote events); URIs have allowed to link them to external resources (for example controlled vocabularies).

A set of controlled vocabularies have been created and integrated with the metadata schema in order to enable the effective integration of the resources.

The following others vocabularies have also been integrated with the LIPARM schema:

- persons
- roles and offices
- chronologies (parliaments/sessions/sittings)
- proceedings, legislative and non-legislative
- constituencies

The list of the produced vocabularies comprises:

1. [Calendar Object Types](#) (e.g. parliament, sitting) ([RDF](#))
2. [Functions and roles](#) (e.g. Prime Minister) ([RDF](#))
3. [Legislative Stages](#) ([RDF](#))
4. [Proceedings Group Types](#) (e.g. Act of Parliament) ([RDF](#))
5. [Proceedings Object Types](#) (e.g. Act of Parliament) ([RDF](#))
6. [Stormont Acts](#) ([RDF](#))
7. [Stormont Constituencies](#) ([RDF](#))
8. [Stormont Members](#) ([RDF](#))
9. [Stormont Parliaments](#) ([RDF](#))
10. [Stormont Sessions](#) ([RDF](#))
11. [Westminster Acts](#) ([RDF](#))

---

<sup>7</sup> For information about LIPARM, see also <http://www.ariadne.ac.uk/issue70/gartner>.

12. [Westminster Bills \(RDF\)](#)
13. [Westminster Constituencies \(RDF\)](#)
14. [Westminster Members \(RDF\)](#)
15. [Westminster Parliaments \(RDF\)](#)
16. [Westminster Private and Local Acts \(RDF\)](#)
17. [Westminster Sessions \(RDF\)](#)

Each item has been assigned a unique URI by which they can be referenced from within the parliamentary metadata schema or any other source (for example the Parliamentary proceedings encoded in [TEI](#)).

MADS ([Metadata Authority Description Schema](#)) XML standard has been used to encode the controlled vocabularies.

Vocabularies have been published as MADS XML files and as [MADS-RDF](#) OWL ontology; they are also available via a web service.

## 2.2 Aggregation Approach

The so-called “aggregation projects” bring content together in a unique point of access through which users are able to reach and access resources otherwise scattered among myriads of sites and microsites. Single providers or groups of providers supply their contents that in turn are processed and integrated in a unique infrastructure according to a predefined schema.

In this framework, approaches and methodologies used for the construction of the controlled vocabularies can be different depending on the types of data supplied by the providers; the knowledge domain they are dealing with; the quality of the already available vocabularies; and the aims of the projects themselves.

### 2.2.1 ArchivesHub

ArchivesHub<sup>8</sup> is a UK based project that encourages archival institutions to add to their EAD exports access points of the following categories:

- people
- families
- organisations
- places
- subjects

Less-used keywords are titles, genre and function.

CHIs should specify the rules and source used. The recommendation is to use NCA Rules (UK National Council on Archives rules) or sometimes AACR2 for subjects. For the entities, the recommended thesauri are UNESCO, LCSH and UKAT (the UK Archival Thesaurus).

---

<sup>8</sup> <http://archiveshub.ac.uk/>

At the moment, the project does not perform normalization of authorities. For instance, if one uses the search for keywords and gives as search parameter “Wellington”, one will get all the possible realizations of the entity “Arthur Wellesley, Duke of Wellington” as different keywords. The project is currently moving towards implementing a system to handle authorities that will allow to cluster their different linguistic realisations<sup>9</sup>.

ArchivesHub is implementing a new EAD editor to be provided to the data providers with VIAF look-up function. Although not all names used to index archival material are in VIAF, the alignment of some of them will contribute to a higher consistency and interlinking of the material<sup>10</sup>.

### 2.2.2 Israel Archives Network

The Israel Archives Network (IAN)<sup>11</sup> aims to integrate archival collections from the various archives in the country. The project integrates metadata from all description levels, as well as digital objects. The focus of the project is the cultural and historical heritage of Israel.

The project provides the CHIs with tools<sup>12</sup> for the edition of EADs which include vocabularies of the National Library of Israel (NLI) and a subset of the LCSH translated to Hebrew<sup>13</sup>. In addition, CHIs can load their own controlled vocabularies into the editor to add links to their own access points.

Description units from different institutions are interlinked using the thesauri of the NLI. Currently, the keywords contributed by the CHIs are not used for interlinking purposes, and are only indexed for free text search.

The IAN aims to implement a crowdsourcing platform to add keywords and subjects to collections that have been poorly described<sup>14</sup>.

### 2.2.3 CENDARI

CENDARI is a project for digital historical research. Its aim is to make the identification of resources useful to historical research easier, focusing especially on existing digital finding aids and assets. It started from two pilot sub-projects concerning medieval cultural heritage and the First World War.

CENDARI developed a metadata strategy combined with domain ontologies. The CENDARI metadata strategy is based on the conceptualization of data related to institutions and their archives at three levels:

- Institutional level

---

<sup>9</sup> Jane Stevenson, ArchivesHub Service Manager. Personal communication.

<sup>10</sup> Jane Stevenson, ArchivesHub Service Manager. Personal communication.

<sup>11</sup> <http://web.nli.org.il/sites/NLI/English/library/aboutus/now/projects/IAN/Pages/default.aspx>

<sup>12</sup> Brief description of the tools:

<http://web.nli.org.il/sites/NLI/English/library/aboutus/now/projects/IAN/standards/Pages/tools.aspx>

<sup>13</sup> Chezkie Kasnett, IAN, Digital Projects Manager. Personal communication.

<sup>14</sup> Chezkie Kasnett, IAN, Digital Projects Manager. Personal communication.

- Collection level
- Item level

For each of these levels new and already existing metadata schemas have been adopted (integrated metadata strategy):

- EAG ([Encoded Archival Guide](#)) for Holding institutions
- CCS ([CENDARI Collection Schema](#)) for Collections
- [MODS](#) (Metadata Object Description Schema, extended with TEI and CCS) for Items.

Part of the CENDARI metadata strategy is the integration of ontologies; very often, in fact, metadata elements, as well as the contents of these elements, are defined concepts. Ontologies are integrated with the metadata records in order to facilitate discovery and linking of resources scattered among institutions.

CENDARI developed an “*ontology system*” that includes small controlled vocabularies as well as highly integrated domain ontologies whose concepts, entities and reciprocal relationships are associated with archives.

Ontologies in CENDARI:

- metadata schemas (see above)
- controlled vocabularies associated with metadata records: lacunae causes; certainty of dates; role of person (associated with collection); material type, etc..
- ontology of sources (type of sources);
- domain ontologies where concepts, entities and reciprocal relationships are associated with the two areas of research (WWI, medieval heritage).

In the case of the two areas of research, ontologies have been preferred to traditional controlled vocabularies because they include a richer set of relationships.

If the authority lists provide authorized names of persons and places, and the thesaurus provides the hierarchy of the relationships between the places, ontologies are able to include the relationships between persons and places, and between persons themselves (parental relationships)<sup>15</sup>.

## 2.2.4 Getty Vocabularies

The Getty Vocabulary Program (GVP, <http://www.getty.edu/research/tools/vocabularies/>) is a long-term (at least 25 years) thesaurus development effort undertaken by the Getty Research institute, part of the Getty Trust. It develops and maintains the following thesauri, at least the first three of which have become authoritative in the cultural heritage domain:

- Art and Architecture Thesaurus (AAT): 45k concepts from visual arts and architecture, including materials, periods, styles, object types, etc.
- Thesaurus of Geographic Names (TGN): 1.2M places
- Union List of Artist Names (ULAN): 280k records of artists, patrons and other people

---

<sup>15</sup> See CENDARI - [Guidelines for Ontology Building](#),

- Iconography Authority (IA): commonly occurring subjects for works of art, including religious, mythological and literary persons, mythical and literary places, imagined or real events, etc.
- Cultural Objects Name Authority (CONA): works of art, series, collections

The thesauri share a common core structure (facets, hierarchies, terms, languages, sources, contributors, etc.). They are structured according to ANSI and ISO standards on thesaurus construction. AAT in particular is praised for its good hierarchical structure, which is based on the following principles:

- 7 fundamental facets for different kinds of concepts
- Guide terms, which are organizational nodes in the hierarchy (not concepts) serving to introduce a level of concepts based on a common distinction, e.g. <chairs by function> (with children Child seat and Throne) vs <chairs by form> (with children High-chair vs Chaise-lounge)

The GVP vocabularies have been published as LOD by Getty in partnership with Ontotext (<http://vocab.getty.edu>). The first 3 vocabularies are in production, and the last 2 are in development. The semantic representation

- is documented extensively: <http://vocab.getty.edu/doc>;
- uses a number of external ontologies ([http://vocab.getty.edu/doc/#External\\_Ontologies](http://vocab.getty.edu/doc/#External_Ontologies)):
  - SKOS, SKOSXL, ISO 25964 for representing thesaurus information. In particular, SKOSXL allows provenance and other information to be recorded about labels, and the latest standard on thesauri (ISO 25964) allows Guide Terms to be modeled as iso:ThesaurusArray
  - DC, DCT for common properties
  - BIBO, FOAF for sources and contributors
  - WGS, Schema for geographic information
  - Bio, Schema for agent information
  - PROV for revision history
  - RDF, RDFS, OWL, XSD for system properties;
- uses a specially developed ontology for custom classes and associative relations (<http://vocab.getty.edu/ontology>)
- is published according to accepted LOD principles and served in a variety of formats ([http://vocab.getty.edu/doc/#Semantic\\_Resolution](http://vocab.getty.edu/doc/#Semantic_Resolution));
- provides a SPARQL endpoint, a large number of sample queries (<http://vocab.getty.edu/doc/queries>) and a special sample query interface (<http://vocab.getty.edu/queries>).

The GVP LOD publication has received acclaim from the CH community and is widely used in scenarios such as:

- Online lookup of terms from museum collection management systems and from search interfaces, using FTS and auto-completion (see [http://vocab.getty.edu/doc/queries/#Full\\_Text\\_Search\\_Query](http://vocab.getty.edu/doc/queries/#Full_Text_Search_Query))
- Pivot vocabulary for coreferencing other thesauri in the domain



The Getty vocabularies grow constantly through contributions. The vocabularies depend on each other (roughly in the order they are listed above). So when an artwork record is received for aggregation in CONA, before it can be ingested all terms appearing in it need to be checked (coreferenced) against the other thesauri, and perhaps a new concept or label added:

- Artwork types, materials, event types are checked against AAT
- Places of production, repository (owner) etc. are checked against TGN
- Artists, architects, sponsors etc. are checked against ULAN
- Subjects are checked against IA
- Finally, CONA artworks can refer to other CONA artworks (be that as subjects, or associative relations like “copy of” or “study for”)

The other thesauri also refer to lower-level thesauri; e.g. ULAN refers to AAT for person roles and event types, and to TGN for event places.

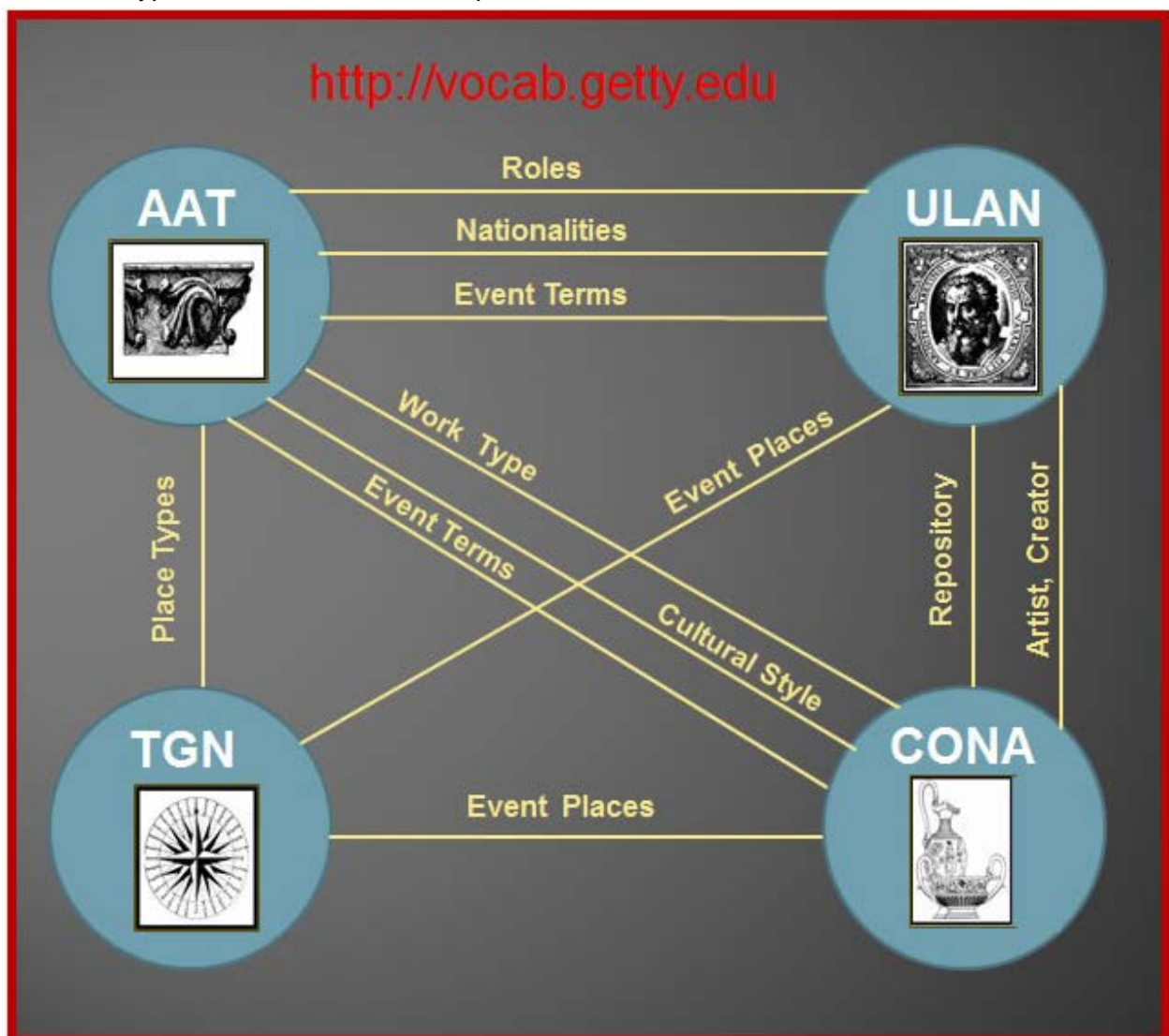


Figure 1 The Getty vocabularies. Credit: Joan Cobb, Getty Research Institute

## 2.3 Conclusions

In addition to the projects described above, there are other projects that could be mentioned here and that could provide ideas for developing the EHRI vocabularies strategy: [APEX](#)



(Archives Portal Europe network of excellence), [Athena](#), [Researchspace](#), [ARIADNE](#), and the [SNAC program](#) (Social Networks and Archival Context, a cooperative program for maintaining information about people documented in the collections).

All projects mentioned here are metadata integration projects - regardless of the topics they deal with or the adopted approaches for the vocabularies.

It is important to highlight that the preliminary definition of a metadata schema or data-model highly affects the identification of the controlled vocabularies.

As is shown in some of the detailed examples, the metadata schema provides the basic starting point for the vocabulary strategies.

The metadata schema is usually the real "pulsating heart" of the projects, that is to say, the "place" where the integration really occurs (and where everything must also be reconducted).

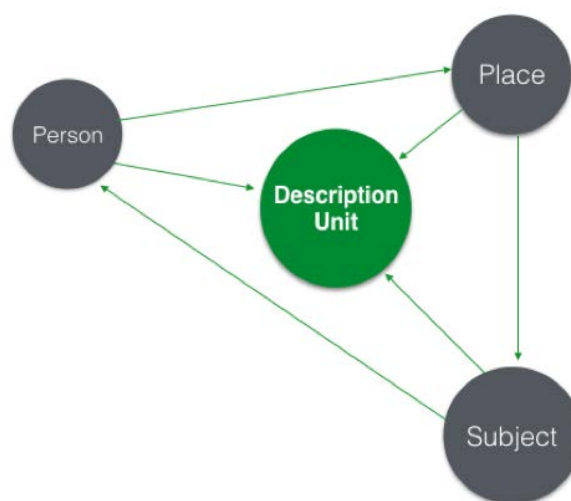
The controlled vocabularies serve to make explicit and to specify the details of some of the key components of the metadata schema, especially those identified as useful access "channels" to the contents. Recurring vocabularies are persons, places, subjects. Then, according to the specific topic of the project, other vocabularies are established.

Vocabularies represent the "arteries" by which it is possible to reach the contents described through the metadata schema.

For instance, if the "heart" consists of the Collection, as it is in the case of the CENDARI project, or the Description Unit, as in the case of EHRI, the vocabularies are modelled in order to reflect the peculiarities of the content descriptions, and to get as much as possible from the retrieval information system.

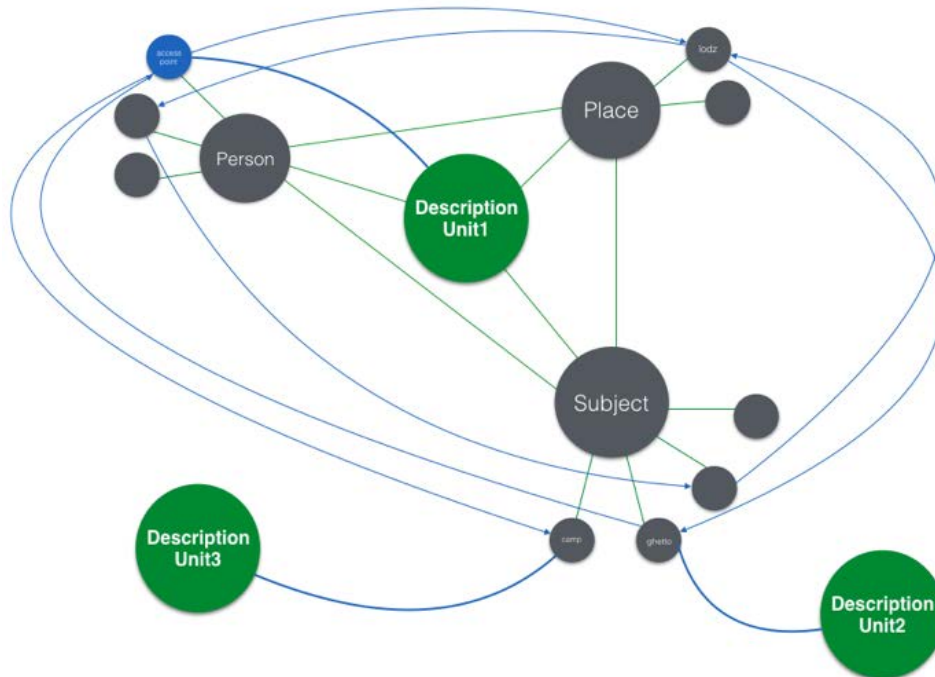
Similar to the circulatory system, the vocabularies should be all interconnected. In fact, in many of the above projects, vocabularies are interlinked.

This connection system keeps you from ever losing the way toward the goal - whether it is the collections, the description units, the digital objects, or whatever the core of the project is.



When the vocabularies (i.e., concepts and terms) are connected to each other, the retrieval of content is strengthened: an access point can be the direct gate to a specific content, but it can also be the gate towards other, sometimes even unexpected, contents.

In this way, through a single access point, we are able to get not only the content we are looking for (the description unit, in case of EHRI) but also all the other description units connected to the concepts/terms linked to the starting access point.



The interlinking of the vocabularies creates the circularity of the information, which is a way toward the enrichment of the knowledge.

In this perspective, the Linked Data (and the ontologies, to define the relationships) can be used to create this sort of "circularity" of information. That is what has been done, for instance, by the LIPARM project, as well as CDEC Digital Library and CENDARI projects.

### 2.3.1 Comparison of the two approaches and relationship to EHRI

Both approaches are useful:

- Aggregation aims to create a central authority in a domain, which provides higher value for users because they can formulate more intuitive and more powerful queries against a unified authority, rather than several disparate (even if linked) authorities.
- Interlinking allows several autonomous authorities to be developed by different communities and for different purposes, yet allows connections to be established between them. Additionally, it is less work since getting a large research community to agree on one central authority takes a lot of editorial work and

consensus building. Finally, linking to external data (e.g. DBpedia, VIAF) can bring benefits in terms of saving data entry (e.g. of vital dates/places) and linking to other datasets.

In particular, we can take the following time-proven advice on the EHRI authorities from the Getty vocabularies:

- It takes a great deal of persistence, attention to detail, and effort to build an authoritative thesaurus in a particular domain; but the benefits to the respective research community are also high.
- When a complex object (e.g. CONA artwork or EHRI archival description) is being ingested by an aggregator, the access points used in it need to be checked/ coreferenced against Authorities. Eventually new labels or authority records will need to be added to these authorities. The growth of the authorities using contributions from the community puts the domain objects in context, and creates implicit relations between them that enable conceptual search.

Given the evaluated pros and cons of the two options, and given the impossibility to change the data provided by the CHIs, the interlinking approach seems to be the more feasible for EHRI.

### 3 Existing EHRI controlled vocabularies

This section describes EHRI controlled vocabularies as implemented in the first phase of the project. We describe these vocabularies, present their usage, discuss their usability as a retrieval tool, and summarise detected quality problems. Subsection (3.2) gives then an overview of the Linked Open Data version created from a subset of the vocabularies and discusses necessary improvements.

EHRI 2 vocs. system

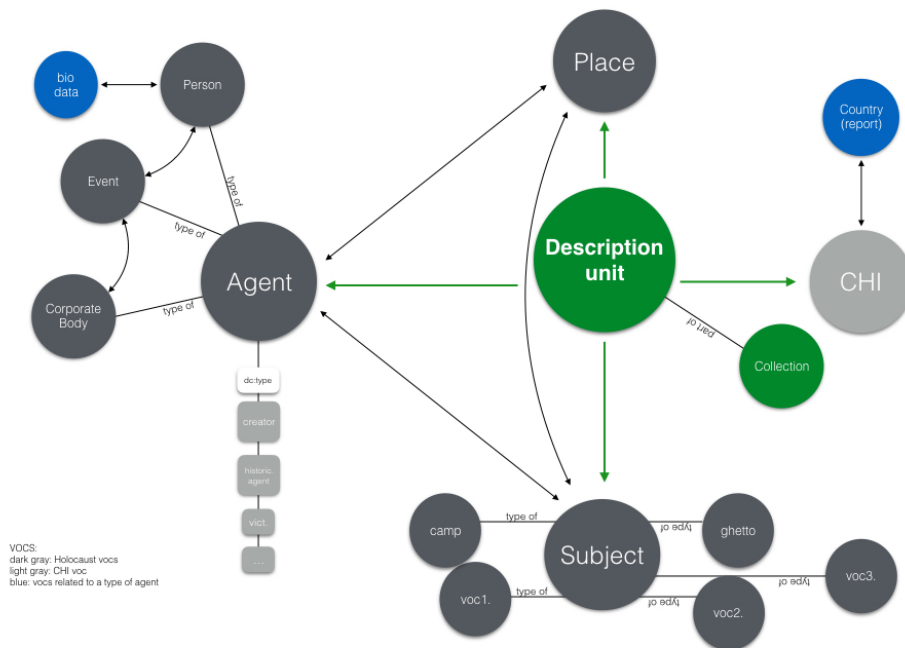


Figure 2: Vocabularies system at EHRI

### 3.1 Vocabularies imported into the portal

#### 3.1.1 EHRI-1 vocabularies

In this subsection we describe the vocabularies provided by EHRI-1 and we provide an overview of the increase of the number of entries in the portal.

##### 3.1.1.1 Description of the vocabularies

The current vocabularies cover, with different degrees of granularity, the following concepts and entity types:

- **Subjects/Concepts:** The main vocabulary is EHRI concepts. We have other keyword lists created with particular purposes, such as to define a narrative of virtual collections, or to interlink material of an institution (NIOD).
- **Person:** as creator, historical person and as Shoah victims (only Terezin ghetto)
- **Corporate Bodies**
- **Ghettos**
- **Camps**
- **Events** (A small set of events. It has not been published yet.)

- **Administrative Districts**
- **Misc.** FAST keywords – contains, along with concepts, other elements, such as geographic features.

### Concepts

The EHRI Thesaurus is a controlled multilingual vocabulary containing 881 concepts linked by hierarchical and associative links, organized according to ISO Standard 25964-1. For each concept there are preferred labels in 10 different languages. Some of them contain alternative labels. Very few have definitions (scope notes).

### Authority List of Person Records

The vocabulary contains 620 personality entries with additional biographical information. The original vocabulary contained only basic information such as first and last name and minimal biographical data. Entries were enriched with keywords from the FAST vocabulary (see 3.1.3), and bibliographical information. Parts of the data, such as the bibliographic information, were not imported into the portal.

### Authority List of Corporate Bodies

The Authority List of Corporate Bodies contains 3,229 entries corresponding to organizations which played a role in the events of the Holocaust. This Authority list includes organizations from 30 January 1933 to the present.

### Authority List of Ghettos

The Authority List of Ghettos consists of 1,109 Ghetto names in English and Hebrew. The list is based on the entries of the Online Encyclopedia of Ghettos of Yad Vashem<sup>16</sup>.

### Authority List of Camps

The Authority List of Nazi Concentration Camps consists of 1,975 Camps' and sub-camps' names. Camps and sub-camps are linked through hierarchical links. The authority list is based on a hierarchical list provided by the International Tracing Service (ITS).

### Authority List of Events

The Authority List of Events consists of 76 events, expressed as concepts and as Simple Event Model (SEM)<sup>17</sup> resources. It has not yet been imported into the portal.

### Authority List of Administrative Districts

Administrative divisions in the German Reich and in the Nazi occupied territories: The time frame is 1933-1945. This authority list of administrative districts includes 261 district records, covering 7 German administrative territories. District records and administrative territories are linked through hierarchical links.

#### 3.1.1.2 Import and manual introduction of data

After the vocabularies were imported into the portal more entities were added by EHRI project partners. We detected a high variation of size in two types of entities, the

---

<sup>16</sup> [http://www.yadvashem.org/yv/he/research/ghettos\\_encyclopedia](http://www.yadvashem.org/yv/he/research/ghettos_encyclopedia)

<sup>17</sup> <http://semanticweb.cs.vu.nl/2009/11/sem/sem.doc.html>

Personalities and the Corporate Bodies. The rise in number of those entity types is motivated by the import of new collections from different geographical areas or related to different historical events to the previously existing data.

Manually introduced data is not part of the Linked Open Data version of the vocabularies.

Vocabulary	Entries created in EHRI-1	Entries in the portal	Percentage of new entries
Personalities	445	620	28.2%
Corporate bodies	715	3,229	77.86%
Concepts	880	881	0.1%
Ghettos	1,106	1,109	0.1%
Camps	2,055	1,975	0%
Administrative Districts	261	261	0%
Events	76		

*Table 1: Entries in the EHRI vocabularies*

### 3.1.2 Research guides vocabularies

In the previous phase of the EHRI project two "Trans-institutional Research Guides" were created, which combine metadata records from different CHIs about two different topics: the Terezin Ghetto and the Jewish Communities during the Second World War. A set of controlled vocabularies was created with the vision of providing multi-faceted access points.

The vocabularies used in the Terezin Guide are:

- **terezin-jewishcouncil:** List of 349 keywords which describes the administrative structure of the "Council of Elders".
- **terezin-keywords:** List of 418 keywords
- **terezin-victims:** List of 8,819 victims
- **terezin-persons:** List of 59 relevant personalities in Terezin.
- **terezin-places:** List of 1663 locations in Terezín. It includes places inside of the city with their GPS coordinates.

Vocabularies of the Jewish Communities guide:

- **jc-persons:** List of 565 person authorities. Around 10% contain relevant biographical information.
- **jc-organisations:** List of 201 organizations. It has been imported as keywords and not as an authority list.

- **jc-places:** List of 1,048 relevant places

### 3.1.3 Additional vocabularies

Two extra vocabularies were created for experimental purposes regarding interlinking of the data:

- **Fast Keywords:** Selection of Keywords of the FAST vocabulary<sup>18</sup> provided by the OCLC<sup>19</sup>. It was created during the enrichment process of the personalities in order to have links between them. After their import they were used by partners of WP15 of EHRI-1 to manually add access points, although some of them corresponded with concepts of the EHRI thesaurus or authorities.
- **NIOD-Trefwoorden:** Selection of the keywords used by the NIOD. The keywords were extracted from the descriptions included in EHRI. NIOD provided mapping between the keywords of the NIOD and the EHRI thesaurus created by WP18, but mapping was not available for every keyword.

### 3.1.4 The EHRI Thesaurus as a retrieval tool

#### 3.1.4.1 Usage of the vocabularies in the portal

The usage of the EHRI-1 controlled vocabularies<sup>20</sup> is summarized in Table 2. In the table we see that there are two types of links:

- Automatically introduced links, where the mapping to the vocabularies has been provided by the CHI before the import in most of the cases<sup>21</sup>
- Manually introduced links created by EHRI to index manually imported descriptions

Vocabulary	Automatic	Manual	Total
Thesaurus concepts	15,773	2,581	18,354
Ghettos	1	276	277
Camps	1	100	101

<sup>18</sup> <http://www.oclc.org/research/themes/data-science/fast.html>

<sup>19</sup> <http://www.oclc.org/>

<sup>20</sup> We include only the EHRI-1 vocabularies in the analysis since the other vocabularies have been created for special purposes.

<sup>21</sup> For one of the datasets, the import of the ARA book, EHRI provided the mappings.

Administrative districts	-	7	7
Persons	1,133	397	1,530
Corporate bodies	-	3,252	3,252
Total	-	3,252	3,252

*Table 2: Links to the EHRI vocabularies in the portal*

From a total of 152,698 descriptions in the portal, only 12,674 are linked to the vocabularies. In the case of automatically introduced links, it does not mean that all the possible mappings between access points provided by the CHI have been mapped to the EHRI thesaurus. For instance, Yad Vashem provided mappings to Thesaurus concepts and Personalities, but not to Places, Ghettos or Corporate Bodies. In a similar way, keywords extracted from the NIOD data were mapped only to the thesaurus concepts, not to authorities.

Only a minority of the description units are linked to the thesaurus, and in most cases the linking is not exhaustive, applying only for some categories of access points. Furthermore, the different vocabularies are disparate (there are no relations connecting concepts from different thesauri). This makes it impossible to find all descriptions referring to the same thing but using different thesauri.

The consequence is that only a minority of description units can be retrieved using the thesaurus, and conceptual (semantic) search is impossible. In the current portal state, only free text search and filtering through the provided facets are available. However:

- This limits findability since it will not find alternative spellings of the same concept, place or person
- It cannot leverage hierarchical relations between concepts.

### **3.1.4.2 Access points provided by the CHIs**

An additional factor that reduces usability of the thesaurus is the fact that CHIs use access points (EAD <controlaccess> elements) that are not part of the EHRI thesaurus. As an example, the EHRI thesaurus has 1,938 concepts distributed in different vocabularies. The ingested collections use 61,771 unique strings as subjectAccess, which are not related to the EHRI thesaurus or interlinked, as in the following example<sup>22</sup>:

wereldoorlog (1939-1945)--joden--redding--frankrijk

We performed a detailed analysis of access points ([Google document](#)). In a first estimation, only 3% of subjectAccess points are formalized as Concepts. This harms discoverability and makes the search less functional than it could be.

The large amount of heterogeneous data ingested into the portal makes it necessary to de-

<sup>22</sup> From description unit <https://portal.ehri-project.eu/units/be-002112-ab-1483>



duplicate the access points and the implementation of data workflows for dynamic enlargement and adaptation of the thesaurus.

### Total access point instances

The following table shows the number of access point instances.

- The vertical dimension shows the subject (item that contains the access point) of which 99.6% are documentary units; and the type of link (of which 53.7% are subjectAccess).
- The horizontal dimension shows the type of access point when it is made as an Authority object. Object=NONE (83%) are access points. for which no corresponding Authority object has yet been created in EHRI.

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%	
<b>Total</b>	<b>295</b>	<b>94847</b>	<b>39</b>	<b>23352</b>	<b>5</b>	<b>567994</b>	<b>686532</b>		<b>17.27%</b>	
Percent	0.04%	13.82%	0.01%	3.40%	0.00%	82.73%				

Table 3: Number of access point instances in the portal

- Red numbers represent errors, e.g. link=corporateBodyAccess should point to object=historicalAgent (or NONE), not to cvocConcept (concepts are not corporate bodies)
- The largest number of access point relations are subjectAccess (53%), placeAccess (23%), personAccess (12%); corporateBodyAccess is also well represented (6.03%).
- creatorAccess (3.9%), which represent creators of archival materials, and historicalAgent (3.4%) are also very important although their relatively small number.

### Unique Access Points (with Type)

The table below, in which we have removed the containing Subject item from the vertical dimension, shows unique access points.

- Unique access point. strings: 141.9k. Each access point is used on average 5.4 times.
- This large number is due to spelling variations and the use of compound (precoordinated) access points. We show how to reduce these numbers in the section 3.1.4.2.4 below.

- 9% of unique access points are made out as **both** string and object (in some case even several objects).

Unique a.p.	object								
Link	country	cvocConcept	documentaryUnit	historicalAgent	repository	NONE	Total	Perc obj	
associative				221			221	100.00%	
corporateBodyAccess		575	21	222	4	4706	5528	14.87%	
creatorAccess		3		2593	1	6784	9381	27.68%	
familyAccess		1	1	5		69	76	9.21%	
genreAccess						440	440	0.00%	
hierarchical				51			51	100.00%	
otherAccess	2	19	3	28		4	56	92.86%	
owl:sameAs						3	3	0.00%	
personAccess				9331		35519	44850	20.80%	
placeAccess	28	2292				15200	17520	13.24%	
subjectAccess		1938	4	6		61771	63719	3.06%	
temporal				41			41	100.00%	
<b>Total</b>	<b>30</b>	<b>4828</b>	<b>29</b>	<b>12498</b>	<b>5</b>	<b>124496</b>	<b>141886</b>	<b>12.26%</b>	
Percent	0.02%	3.40%	0.02%	8.81%	0.00%	87.74%			

Table 4: Unique access points

- Again, red numbers represent invalid combinations. For example, placeAccess should be made into Place authority objects (e.g. coreferenced to Geonames), and not simply cvocConcept (concepts).
- The link type cannot always be trusted. For example, in this combination the first link type is correct but the second is not:
  - AE.G. corporateBodyAccess
  - AE.G. subjectAccess

### 3.1.4.2.1 Unique Access Points (strings)

The following table shows the number of unique access points as strings (ignoring any object type).

Type	Count	Perc
associative	221	0.17%
corporateBodyAccess	5142	3.95%
creatorAccess	8805	6.77%
familyAccess	76	0.06%
genreAccess	440	0.34%
hierarchical	51	0.04%
otherAccess	56	0.04%
owl:sameAs	3	0.00%
personAccess	36736	28.24%
placeAccess	15559	11.96%
subjectAccess	62933	48.39%
temporal	41	0.03%
<b>Grand Total</b>	<b>130063</b>	

Table 5: Unique access points as strings

- 1.6-1.9% of access points appear under two types, e.g. creatorAccess and personAccess or creatorAccess and corporateBodyAccess:
  - A. Χαμπούρης creatorAccess ehri-pers-000544  
historicalAgent
  - A. Χαμπούρης personAccess ehri-pers-000544  
historicalAgent
  - Association of Jewish Refugees in Great Britain  
corporateBodyAccess
  - Association of Jewish Refugees in Great Britain creatorAccess
- Some access points also appear as 2 Authority objects, and also as a mere string. Of course, we should merge all these into a single Authority object.

As mentioned in (3.1.4.1) only a minority of the description units in the portal have their subjects and authority access points linked to the vocabularies of the thesaurus. For the rest of description units, access points are just strings that are not linked to any entity on the portal. Although those access points are indexed by the free text search engine, this indexing is not only unable to resolve the retrieval of information in different languages, but we have detected a high degree of heterogeneity in the data even for description units in the same language. Here we will enumerate the most relevant cases that we found in the analysis of the data imported to the portal.

In the case of other entities (personAccess, corporateAccess, geographicAccess, etc.) the situation is even worse, since the EHRI authorities in these areas are much smaller.

Here we present some examples of problems that we have found in an initial data analysis, and our recommendations for further work.

### 3.1.4.2.2 Compound Access Points

Many access points are compound (pre-coordinated) concepts. Those compounds often group different kinds of authorities, as for instance in:

- School children--Germany--Frankfurt am Main--1930-1940: Subject+Place+Date
- Poland--Economic conditions--20th century: Place+Subject+Date
- Children from Bialystok: its use for search would be more efficient if we split the concept from the place

USHMM clearly indicates compound subjects using the LCSH notation "--" (though the number of dashes varies). Many other CHI use commas or some other notation. In some cases different kinds of information have been concatenated without any composition of the meaning<sup>23</sup>, as in the following example:

- Aachen, Arnsberg, Baden, Bayern, Berlin, Bremen, Düsseldorf ...

<sup>23</sup> The most extreme case that we found is a long list of entities of different types that have been put together in a string: [https://portal.ehri-project.eu/units/nl-003006-easy\\_collection\\_2-3-urn\\_nbn\\_nl\\_ui\\_13\\_ait\\_bgg](https://portal.ehri-project.eu/units/nl-003006-easy_collection_2-3-urn_nbn_nl_ui_13_ait_bgg) This representation corresponds to the original hosted by the CHI: <https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:41264/tab/1>

The link type applies only to the first atom (usually), e.g. in:

- subjectAccess “abortion--Poland--Oswiecim”

(abortion in the Oswiecim camp in Poland), subjectAccess applies only to “Abortion”, the other two atoms are placeAccess. So not only can the link type not always be trusted, but it provides limited guidance in the case of compound access points.

Compound concepts are problematic because they lead to a combinatorial explosion, and cannot be coreferenced to bound Authorities. They should be broken up (atomized), coreferenced to Authorities, and then reapplied in a conjunctive manner. For the example above, we should link to subject authority Abortion, and place authority Oswiecim. Poland, being an ancestor place, can be deduced and does not need to be applied directly. However, it can and should be used for disambiguation of the more specific place.

We do not intend to remove access points from archival descriptions submitted by the CHI: these should be kept as originally received data. We intend to complement them with links to authorities in order to facilitate discoverability through semantic search. Authorities themselves should avoid compound subjects as much as possible.

### 3.1.4.2.3 Spelling Variations

We found that concepts which could have similar meaning are represented by different strings with some small differences, e.g. refugee (singular) vs. refugees (plural).

Different **spellings and conventions** for writing the same entity are used widely by the different CHIs. We have cases like the Lodz ghetto, which is realized by 24 different strings in the portal. The case of Lodz illustrates very well some of the different problems that we found:

- Spelling errors.
- Different spellings in the transliteration of the original Polish word into English. Some transliterations conserve the Polish diacritics (Łódź) and some use only the available characters of the English alphabet (Lodz). Furthermore, different Unicode representations are used (combining characters vs accented characters), i.e. no Unicode normalization is performed.
- Punctuation and Local identifiers: many access points use a variety of punctuation, and/or are prepended by a CHI-specific local identifier that is not useful.
- Different conventions: Different CHIs use different conventions to represent a ghetto
  - Using just the name: Łódź; Lodz, etc.
  - Specifying function: Łódź (Ghetto),
  - Specifying function and location: Lodz ghetto, Poland; Lodz,Ghetto,Poland  
Ghetto represented as place: Lodz, Poland, Eastern Europe;
  - Multilingual entry: Łódź – Łódź – Polen Łódź – Łódź – Poland
- There are cases in which a concept appears in more than one language in the same string, as in: oorlogskinderen / warchildren.

For cases in which the same entity is realized through different strings due to different spellings, representation conventions or errors, our recommendation is to cluster them as different labels of the same concept.

### 3.1.4.2.4 Recommendations

We recommend the following actions to connect access points to Authorities:

- **Decomposition:** break compound subjects like “Abortion--Poland--Oswiecim” into atomic concepts like “Abortion” and “Oswiecim (Poland)”. As suggested above, this is non-trivial because a compound can produce atoms of different kinds (e.g. subjectAccess -> subjectAccess, geographicAccess), and because parent places mentioned in the access point should not be treated as independent atoms, but used only for disambiguation. Improper compounds like “Aachen, Arnsberg, Baden...” are even harder to handle.
- **Normalization:** reduce variations in spelling by Unicode normalization, removing accents, removing punctuation and other parasitic patterns by regular expressions (some CHI include CHI-specific identifiers at the beginning of access points, which are not useful), and person name inversion (“last, first” is equivalent to “first last”). We should be careful to preserve parenthesized years in person access points, because that information is useful for identification purposes.
- **Lower-casing** should also be done, but case is often useful to distinguish named access points (places, persons, corporateBodies) from concepts. Unfortunately, many places in French are written in lower case.
- **Deduplication:** use OpenRefine to cluster access points with minor spelling variations. Each cluster represents a single concept, and the access point variants are different labels for it.

Ontotext has conducted experiments in this direction, which are encouraging:

- The number of deduplicated atomic access point instances is 1,282,978. Compared to the number of compound access point instances (686,532), there are 1.87 atoms per access point on average. But that number is a low-end estimate because the number of compounds is not deduplicated. This shows that compound subjects are widely used, and must be dealt with.
- The number of unique atoms is 105,381. Compared to the number of unique compound access point (130,063), this is a 9% reduction. Clustering of the compounds with OpenRefine (i.e. removing variations in spelling) reduces to 100,951 unique atoms, or 4.2%. This number can be improved significantly if clustering is performed on the atoms, which we plan to do.
- Such deduplicated access points are still not a proper thesaurus, but are the beginning of one:
  - **Typification:** determine the kind of access point. For the leading atom, this is indicated by the type of link (e.g. subjectAccess vs placeAccess, though it cannot be trusted in all cases). For other atoms we can use heuristics and patterns: first try Geonames to check whether it is a place name; then check "First Last (birth-death)" or "Last, First (birth-death)" to see if it is a Person; if neither of the previous checks match, we can assume that it is a Subject. Recognizing corporateBodies will be very hard since there is no recognizable pattern for them, e.g. Florian Geyer  
 subjectAccess  
 must be the [8th SS Cavalry Division Florian Geyer](#) so it should be corporateAccess, but that's very hard to determine automatically.

- **Coreferencing/merging:** search the labels of existing EHRI authorities for a match against some of the labels in the cluster; if found, add the remaining unique labels to the concept and skip the next step.
- **Candidate concepts:** Gather the clusters into a “candidate concepts” thesaurus. Assign URLs to these concepts automatically. Although these concepts are not vetted by the EHRI Thesaurus Editorial Board, they may be useful search points for the user (otherwise the CHI would not have included it as an access point). They can still be used in conceptual search.
- **Re-application:** re-index the archival description unit with the concepts or candidate concepts thus identified (existing or newly added URLs).

### 3.1.5 EHRI Thesaurus Quality

We have not done a comprehensive review of the quality of the EHRI Thesaurus. However, a cursory examination – while cleaning up its RDF representation and converting it to SKOSXL – uncovered some problems.

#### 3.1.5.1 Mixed-up Labels

In the thesaurus terms we found some cases of wrong (mixed-up) labels, e.g. concept 738 mixes the English label “Holy Writings” with the German preferred “Homosexuelle Frau”, “Homosexueller Mann” and alternative label “Lesbe” as one can see in this SKOS sample:

```
<http://data.ehri-project.eu/ehri-skos.rdf#tema-738>
a skos:Concept ;
skos:inScheme <http://data.ehri-project.eu/ehri-skos.rdf#> ;
skos:broader <http://data.ehri-project.eu/ehri-skos.rdf#tema-724> ;
skos:related <http://data.ehri-project.eu/ehri-skos.rdf#tema-760> ;
dct:created "2012-08-16 08:13:54" ;
skos-ehri:altMaleLabel "Schwule Frau"@de ;
skos-ehri:prefFemaleLabel "Homosexuelle Frau"@de ;
skos-ehri:prefMaleLabel "Homosexueller Mann"@de ;
skos-ehri:prefNeuterLabel "Schwuler Mann"@de ;
skos:altLabel "Ecriture Sacrée"@fr ;
skos:altLabel "Sacred writings"@en , "Holy scriptures"@en ;
skos:altLabel "Svâtoe Pisanie"@ru-latn ;
skos:altLabel "Svâšenne Pisannâ"@uk-latn ;
skos:altLabel "Священне Писання"@uk-cyrl ;
skos:altLabel "Lesbe"@de ;
skos:prefLabel "Heilige geschriften"@nl ;
skos:prefLabel "Holy writings"@en ;
skos:prefLabel "pisma święte"@pl ;
skos:prefLabel "svaté spisy"@cs , "Svâte Pis'mo"@uk-latn ;
skos:prefLabel "sveto pismo"@sh-latn ;
skos:prefLabel "svâtoe pisanie"@ru-latn ;
skos:prefLabel "écriture sainte"@fr ;
skos:prefLabel "Святе Письмо"@uk-cyrl , "Szent iratok"@hu ;
skos:prefLabel "Святое Писание"@ru-cyrl .
```

After we investigated the causes of this error, we found that in the workflow there were several problematic steps in which persistence of the identifiers was not kept; translators modified the formats of their Excel tables or even worse, re-interpreted the instructions



making the tables uninterpretable; staff at one institution worked to correct them by hand, referring to language specialists of the institution, etc. A lot of the work was done by hand and without assistance of colleagues with a technical background. That made the quality control very hard, especially when taking in account the multilinguality of the data.

Most of the errors can be repaired using just provenance information. For other cases we are not sure whether they are errors and they will need to be reviewed by content specialists. We are aware that the organization model for the elaboration and management of the control vocabularies in EHRI-1 is not sustainable; even less so if we believe that our thesaurus has to grow in order to be an efficient retrieval tool for new materials.

### **3.1.5.2 Translation Problems**

Some of the translations of the labels of the Thesaurus concepts can be considered as questionable, with potential errors whose possible origin has been described in the previous subsection. Let us look at one of them:

Consider <http://data.ehri-project.eu/ehri-skos.rdf#tema-618>. It has label "internirovannye"@ru-latn (interned persons) which:

- Does not have a corresponding label in ru-cyrl
- Does not quite match the meaning (detained persons)

That shows the necessity for translations, addition of labels, editing, etc. to be made in an environment that is able to keep track of changes, allows collaborative edition and supervision of content, and maintains formal consistency.

### **3.1.5.3 Structural Problems**

One of the prominent structural problems is the existence of problematic concepts in the Terezin thesaurus like:

#Root\_node\_for\_keywords\_old  
#Root\_node\_for\_object\_types

They seem to be temporary nodes created for technical purposes during the development process. Special editorial steps are needed to mark such temporary nodes, and omit them when publishing.

The fact that EHRI uses a number of independent (disconnected) thesauri (EHRI, Terezin, Jewish Communities...) can also be seen as a structural problem, since there is not any connection between potentially related or similar concepts. We have for instance three different entries of the Lodz ghetto in different vocabularies:

- <https://portal.ehri-project.eu/keywords/ehri-ghettos-513> in EHRI-ghettos
- <https://portal.ehri-project.eu/keywords/terezin-places-place-iti-48> in Terezin-places
- <https://portal.ehri-project.eu/keywords/jc-places-place-iti-48> in the Jewish Councils places

This provides less value to the users, since by searching for one of these keywords, they cannot find archival descriptions using the other keyword, even though it means the same thing.

### 3.1.5.4 *Lack of scope notes*

A scope note is used<sup>24</sup>:

- to restrict or expand the application of a concept,
- to distinguish between concepts that have overlapping meanings in natural language, or
- to provide other advice on concepts usage to either the indexer or the searcher.

In the EHRI portal where, during data integration, access points of incoming data can be mapped to concepts in the EHRI thesaurus in order to increase interlinking between documents, the correct interpretation of the meaning of the concepts is crucial. The same word (e.g. Transport) could mean different concepts for different Holocaust researchers.

In the current implementation of the thesaurus concepts, only 42 concepts of the thesaurus concepts have scope notes.

### 3.1.5.5 *Recommendations*

The first priority is to solve the problem of mixed up labels using the original tables and rebuilding the SKOS file. After that, the rest of the potential problems related to the content and translations or the edition of scope notes can be done only by an editorial board of specialists in the area.

## 3.2 **Vocabularies published as Linked Open Data**

Not all the vocabularies of EHRI have been published in the first phase as linked data due to two reasons:

1. The Corporate Bodies dataset was not represented in RDF because the initial consideration was that EAC-CPF fulfils the criteria of reusable standardized representation.
2. The published subsets are based only on the initial imports in the first phase of the project, and the entries introduced manually after the imports were not included.
3. Some vocabularies were created just for concrete purposes, such as the FAST keywords or the vocabularies for the virtual collections.

To build the RDF representation of the vocabularies EHRI used the following ontologies:

- Thesaurus concepts: SKOS
- Camps: SKOS
- Personalities:
  - Dublin Core
  - EAC-CPF
  - FOAF
  - RDF
  - VIAF
- Ghettos: SKOS, Geo

---

<sup>24</sup> ANSI/NISO Z39.19 - Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies



- Events: SEM

EHRI created a SKOS extension for languages that has different labels for male, female and neuter instances of a concept (e.g. actor vs. actress). The purpose was to satisfy the SKOS recommended constraint that a concept should have one `prefLabel` per language, not multiple labels. The SKOS extension adds the following labels as a subproperty of `rdfs:label`:

- `skos-ehri:prefMaleLabel`
- `skos-ehri:prefFemaleLabel`
- `skos-ehri:prefNeuterLabel`
- `skos-ehri:altMaleLabel`
- `skos-ehri:altFemaleLabel`

The EHRI vocabularies are not interlinked, or are linked only at the very top level (for instance the top level of the Camps refers to the concept “Camps” of the Thesaurus concepts).

### 3.2.1 Problems detected in the LOD set

Here we present a summary of some of the problems found in the current LOD set.

#### 3.2.1.1 RDF validation problems

The RDF files of the EHRI Thesaurus were validated using Skosify<sup>25</sup>. After checking the validity with other tools, we realized that

- Jena RIOT does not find problems
- Sesame RIO finds validation problems

#### 3.2.1.2 Syntax of the URL

The entries of the EHRI LOD vocabulary sets are defined by URLs like the following example: <http://data.ehri-project.eu/ehri-ghettos.rdf#30>. This violates the following COOLURIs<sup>26</sup> principles:

- The function of the URL is to identify the resource or entity, but not a concrete technical implementation or encoding (RDF/XML in the given example). The URL enables to get data in different languages and data formats using HTTP Content Negotiation<sup>27</sup>.
- Hash URLs are not recommended for (potentially) large sets of data. When a client wants to retrieve a hash URI, then the HTTP protocol requires the fragment part to be stripped off before requesting the URI from the server and the server will return the whole set<sup>28</sup>.

The correction of these errors does not involve a lot of resources but is important for the usability of the data.

---

<sup>25</sup> <https://www.w3.org/2001/sw/wiki/Skosify>

<sup>26</sup> <http://www.w3.org/TR/cooluris>

<sup>27</sup> <http://www.w3.org/TR/cooluris/#conneg>

<sup>28</sup> <http://www.w3.org/TR/cooluris/#hashuri>

### 3.2.1.3 Ontological Modelling

The main ontological problem is the confusion between concept and real world thing in the thesaurus. We illustrate this in the following example:

```
<http://data.ehri-project.eu/ehri-camps.rdf#1141> a skos:Concept ;
  skos:prefLabel "Wien - Magdalenenhof"@de-latn ;
  dc:date "2014-03-31"^^xsd:date ;
  dc:creator "the International Tracing Service (ITS) in Bad Arolsen"@en-
latn ;
  dc:contributor "European Holocaust Research Infrastructure (EHRI)"@en-
latn .
```

If we paraphrase the meaning of this vocabulary entry in English, we read that the Concentration Camp Magdalenenhof was created in 2014, that it was created by the ITS and the EHRI project contributed to its creation, which is of course not the intended meaning.

A solution for this problem has been proposed by the Getty vocabularies: the use of separate URLs for the concept and the real world object<sup>29</sup>. We intend to explore this solution for EHRI.

Other problems in the modelling are:

- Geo-spatial information:
  - Geo-coordinates of the Ghettos are given using the property `geo:lat_long`, which is already deprecated and probably will be removed from the geo ontology<sup>30</sup>.
  - The type `geo:Point` is incorrect: a Ghetto has a geographic shape whose center is a `geo:Point`, but a Ghetto itself is not a point. It is a `geo:SpatialThing`, which is implied by the domains of `geo:lat` and `geo:long`, and allows it to have coordinates
- Language tags: The data includes some unnecessary script codes, as "de-Latn" and "en-Latn". Since Latn is the assumed script for these languages, there is no reason to add it.

### 3.2.1.4 Problems of the EHRI Skos extension

The EHRI Skos extension created some new problems in the thesaurus, such as the lack of preferred labels, and the creation of extra alternative labels without the necessity for them. The most important problems are:

- Since there is no constraint to add more than one alternative label, the proposed `ehri-skos:altMaleLabel` and `ehri-skos:altFemaleLabel` do not have any practical justification. An additional problem is the lack of consistency in their use, since we find alternative labels with and without gender specific label.
- `ehri-skos:prefMaleLabel`, `ehri-skos:prefFemaleLabel`, `ehri-skos:prefNeuterLabel` are all sub-properties of `rdf:label`. This leaves the concept without any preferred labels. The purpose of the unique preferred label constraint is to allow applications to display an

<sup>29</sup> [http://vocab.getty.edu/doc/#Concept\\_vs\\_Thing\\_Duality](http://vocab.getty.edu/doc/#Concept_vs_Thing_Duality)

<sup>30</sup> <http://www.w3.org/2003/01/geo/#vocabulary>

unambiguous label for each concept in each language. Having not preferred labels makes that impossible. That has consequences for the display in the ERHI portal<sup>31</sup>.

- The ehri-skos labels are both `rdf:Property` and `owl:AnnotationProperty`. This can cause reasoning problems in some situations. They should be `owl:DataProperty`.
- The ehri-skos labels do not have any range defined. A range of `rdf:langString` should be used.

---

<sup>31</sup> Data import in the portal follow the SKOS standards, ehri-skos labels are not displayed.  
<https://portal.ehri-project.eu/keywords/ehri-skos-tema-1215>

## 4 Structure of the EHRI vocabulary "universe"

As it has been described above in section 3.1.1, the current EHRI “vocabularies system” is made up of a series of authority lists. Some of them are the result of extractions from the archival descriptions (persons, corporate bodies), others have been imported (FAST Keywords; NIOD Treefwoorden); others have been built for specific purposes, taking concepts from already existing structured sources (Camps, Ghettos); yet others have been built from scratch to describe special collections (Terezin, Jewish Council).

All of these authority lists are structured in their own way and exist separate from each other. For this reason it is quite normal to find duplicated or overlapping concepts and terms; as it is common to find the same concept expressed in different terms.

Given this situation, how can we achieve a structured system that will best cover the Holocaust knowledge domain?

This issue is fundamental when we consider the controlled vocabulary system as a key step towards:

- effective and precise retrieval of contents
- a facilitated indexing process
- archival descriptions provided by (interrelated) points of access

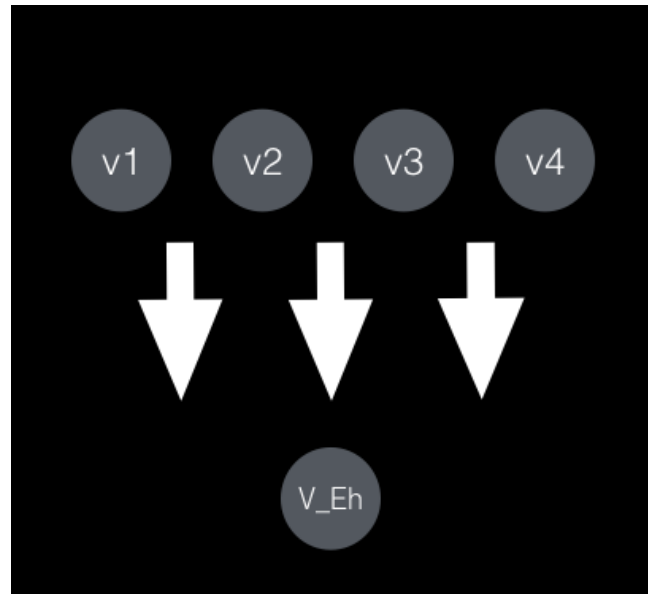
Besides these practical advantages, there is also a wider, long-term aim to be considered in building the EHRI controlled vocabularies: it should represent (and it should be used as) the authoritative thesaurus for describing material related to the Holocaust domain.

Especially for this last reason the EHRI thesaurus should be re-usable for all of the institutions and people involved in the description and indexing of Holocaust resources.

During the WP11 meeting held in Berlin in December 2015, two different approaches to the development of the EHRI thesaurus were discussed: the centralized approach and the federated approach.

### 4.1 Centralized approach

A centralized thesaurus involves the merging of the existing lists/vocabularies into a unique and hierarchically organized structure. The centralized approach implies switching from a multitude of vocabularies to a unique vocabulary.

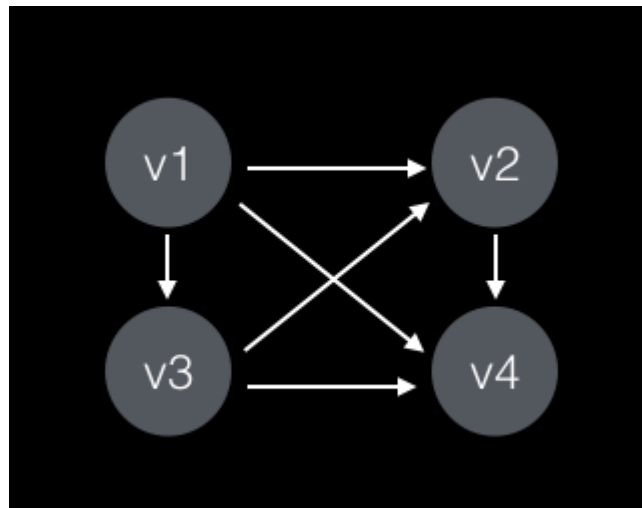


*Figure 3: the centralized approach*

Examples of centralized thesauri are those from The Getty Research Institute (ULAN, TGN, AAT, CONA)<sup>32</sup> as well as the USC [Shoah Foundation Institute Thesaurus](#).

## 4.2 Federated approach

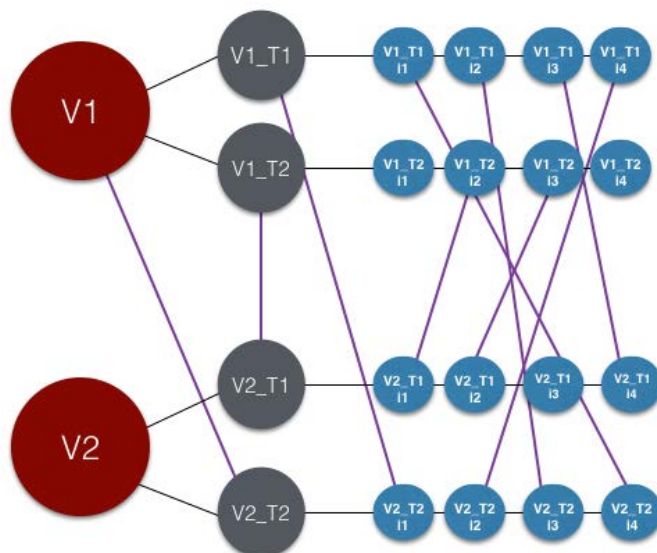
The “federated” approach implies the persistence of all the already existing vocabularies; through mapping they are put in relation to each other.



*Figure 4: The federated approach*

<sup>32</sup> See a thorough presentation of TGN, here [https://www.getty.edu/research/tools/vocabularies/tgn\\_in\\_depth.pdf](https://www.getty.edu/research/tools/vocabularies/tgn_in_depth.pdf) ).

“Mapping” makes it possible to maintain the already existing vocabularies and find correspondences between concepts and terms at all levels between all the vocabularies.



*Figure 5: Mappings between vocabularies*

The user will have the possibility of choosing among a series of vocabularies to search for the concept/term that he/she is looking for (or that best fits the object that is going to be described, in case of editor of descriptions/indexer).

The ingestion of new vocabularies will require new mappings.

A good explanation of the mapping option is provided by GRISP - General Research Insight in Scientific and Technical Publications<sup>33</sup>.

### 4.3 Combinatorial approach

A third option that can be taken into consideration is the integration of vocabularies (combinatorial approach). It is a compromise solution between merging (centralized approach) and mapping (federated approach): high levels (taxonomy) are merged; the low levels (item) are mapped.

<sup>33</sup> GRISP (General Research Insight in Scientific and technical Publications, [Lopez and Romary, 2010]) is a work in progress aimed to create a multilingual terminological database covering multiple technical and scientific fields from various open resources. Its main goal of the database is to support automatic text processing applications. For the present multi-domain terminology GRISP uses 76 basic domains derived from the technical and scientific domains of the lexical database WordNet, and organizes them into hierarchies of concepts. Together with Wordnet other lexica and ontologies (domain specific or general purpose) are used and integrated. The process of integration consists of mapping of concepts based in a set of rules and machine learning techniques. (<https://hal.inria.fr/inria-00490312/document>)

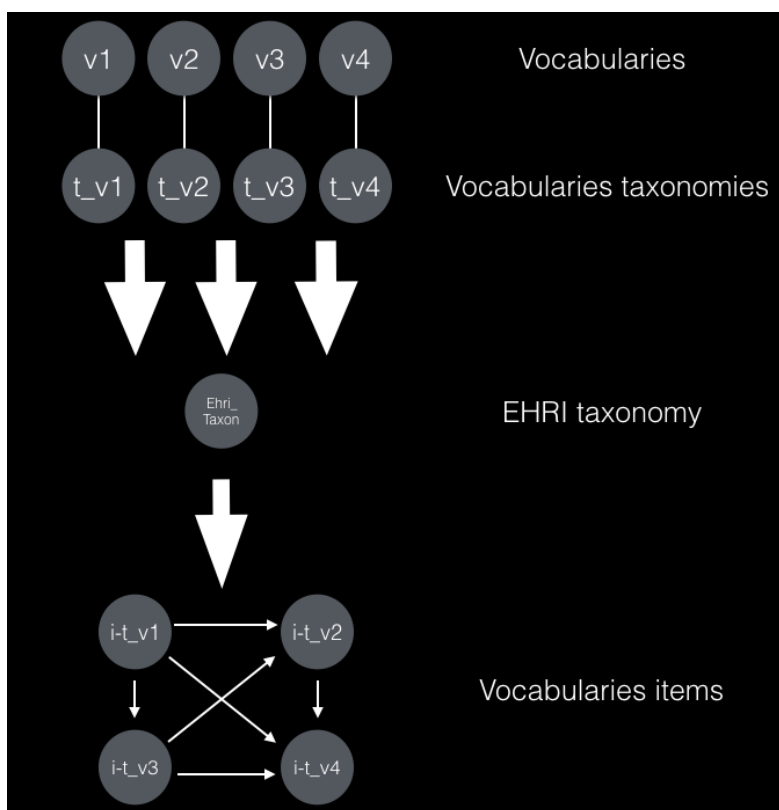


Figure 6: Combinatorial approach

The combinatorial approach involves the structures, not the single items. Through the integration, while the top-nodes are unique and defined, the single items are mapped; in this way they can complement each other. A term, in fact, can vary its meaning on the basis of its provenance, therefore it is important to preserve all its possible meanings (on the other hand, different terms can express the same or complementary meanings). In spite of these different meanings, terms can depend on the same upper-node (taxonomy).

For example:

Persecution (Taxonomy)

- transport (term in Voc 1)
- convoy (term in Voc 2)

In the case of EHRI, thanks to the work done in EHRI-1, this combinatorial approach would seem to be already in practice: while the top-levels (taxonomies) could be provided by the vocabularies mentioned above (see 3.1, maybe they could be increased, if necessary) the low levels will be mapped. This approach seems to be suitable in case of ingestion of new vocabularies coming from new providers.<sup>34</sup>

<sup>34</sup> A case study about this approach is explained in Golden, Shaw, Buckland; *Decentralized Coordination of Controlled Vocabulary*, Conference paper in Proceedings of the American Society for Information Science and Technology, January 2014. Retrieved from <https://www.researchgate.net/publication/275414322>

## 4.4 Conclusions

Even though we preferred the centralized approach during our initial discussion, upon further deliberation it became clear that the federated or combinatorial approaches are more doable in the time-frame of the EHRI-2 project. The retrieval of information in a system adopting a centralized thesaurus is, in general, very consistent and effective, as the Getty's experience clearly shows.

On the other hand, the construction of a centralized thesaurus is a demanding task; so much so that it could be considered the object of a dedicated project. From a practical point of view, the top-down approach implicit in its building (already attempted during EHRI-1), necessarily implies a very substantial contribution of an editorial board and no resources have been explicitly reserved for staffing such a board. Furthermore, the decentralized structure typical of EHRI - and more generally of aggregation projects similar to EHRI - renders the imposition of a new indexing system to partners that often already have their own system in place difficult, if not outright impossible. Finally the centralized approach would involve the alteration of description provided by CHIs which runs against the established data integration practice in EHRI.

Given the vocabularies already ingested and/or set up during EHRI-1, the federated, bottom-up approach appears to be one of the most doable ways for building the EHRI thesaurus. The mapping of already existing terms at all levels (e.g. using `skos:exactMatch`) avoids the loss of meaning.

The main benefit of the federated option is that there is nothing else to do but the mapping of the terms. At the same time some drawbacks have to be considered:

- When indexing, the indexer must not only select the term(s) to use but also the vocabulary; he/she will have to evaluate the implicit or explicit meaning of a term according to its provenance.
- The portal search would have to implement a query expansion of a set of concepts linked by `skos:exactMatch` using several thesaurus trees. Otherwise the portal user would have to select not only the concept but also the thesaurus.
- The mapping of terms with same names but different meanings could generate an improper indexing.
- Finally, the portal will need to implement a very friendly graphic interface for the information retrieval.

The combinatorial approach, which combines the two approaches mentioned above, seems to be the most suitable for EHRI. It is important to note that, given the work done in EHRI-1 and during the first months of EHRI-2, the development of an EHRI thesaurus according to the combinatorial approach is already a work-in-progress. A further analysis of the existing vocabularies is recommended in order to identify other possible topic-nodes (taxonomies). The identified taxonomies could be the real new starting point for a bottom-up approach (the most viable one) and a mapping of the terms. It is also worth evaluating if the current existing vocabularies such as Terezin or the Jewish Council, could be "converted" to topic-nodes.



## 5 Dataflows

The current EHRI Authorities are fragmented, the coverage of archival descriptions is low, and the depth of Authority data is low. To improve this situation we need to do considerable work on various parts of the system:

- Deploying new components, e.g. VocBench for Concept thesaurus management
- Writing new components (e.g. access point deduplication and coreferencing to authorities)
- Extending existing components (e.g. administrative interface of the portal for editing EAC information about agents)
- Adding new publication interfaces (e.g. semantic publishing of authorities)

All these changes will require very careful re-architecting of the respective subsystems and implementing appropriate data flows and synchronization. Some examples:

- When we ingest an EAD (archival description), how do we:
  - match its controlled access points to the EHRI authorities
  - generate candidates/suggestions for the EHRI thesaurus: both concepts and terms in new languages
  - re-apply authority links to the EAD
- When we reingest an EAD, how do we implement an update, i.e. deleting the old data in neo4j and adding the new data, including "garbage collection" of candidate concepts and labels that are not used in another EAD.
- If we decide to edit a specific Authority file not using the administrative interface of the portal, but in a specialized tool (e.g. VocBench for Concepts, Wikidata for Camps/Ghettos), how do we implement data synchronization to the main data store (neo4j).

## 6 Thesaurus Editorial Policies

The amount of EHRI Authorities will increase in the course of EHRI-2 and the structure of some vocabularies may be altered in different ways such as:

- Partial or total merging of vocabularies
- Addition of new concepts in the concept hierarchy, either extracted from ingested material or proposed by specialists
- Addition of new authorities extracted from access points, either extracted from ingested material or proposed by specialists

Those changes in structure and especially the expansion are crucial processes if the Thesaurus will be a good retrieval tool for the content of the portal. We have to ensure that in the process good new entries are selected, incorporated translations and alternative names have a good quality and that the hierarchy is maintained and remodelled in a controlled way. This validation process requires a high degree of domain knowledge and has to be done by an editorial board of specialists in the area.

We recommend that the EHRI Thesaurus should grow through the addition of relevant Access Points from ingested EADs (similar to the way the Getty vocabularies grow, see (2.2.4)):

- Before a description is ingested, all its subjectAccess points are analyzed and, if possible, coreferenced to existing thesauri. This service will be developed in cooperation between T11.4 and T10.3.
- If not found, the subjectAccess is added to a candidate list
- The editorial board periodically works through the candidate list, adds a new label to an existing concept, or recognizes the label as a new concept and puts it in an appropriate hierarchy

This does not mean the EHRI thesaurus should grow to say 5,000 concepts. Maybe it should hold only the 1,000 concepts that are most salient to the Holocaust domain. However, a secondary and wider thesaurus should be more inclusive and include all subjectAccess points that the editorial board can process. This will enable conceptual search and improve discoverability.

### 6.1 EHRI Thesaurus Editorial Board and Workflows

We recommend to the EHRI PMB to establish an editorial board that will be responsible for the evolution and quality of the EHRI thesaurus (or thesauri). An important task is to establish editorial policies and workflows, i.e. the way of working of the editorial board. The following questions should be decided:

- Who will edit the thesaurus?
- What editorial roles are needed?
- What editorial workflows should be established about Concepts and about Labels?
- Who accepts/rejects candidates?
- Are there specialists per language? Do we need to track, for example, requests for translation?
- How do we keep the hierarchy meaningful? (if disorganized, it will collapse on itself)

- How do we make the hierarchy reasonably deep, to facilitate better retrieval through Concept expansion? (When a user searches for a concept, he should find all documents indexed using its descendants.)
- Merging of vocabularies
- Tracking the provenance of concepts and labels, i.e. creation and modification

## 7 Transformation and curation of vocabularies

### 7.1 Transformation and Curation of existing vocabularies

#### 7.1.1 Concepts (Subjects)

Main data requirements:

- Ability to enter multilingual labels in a number of languages (e.g. uk “Ukrainian in Cyrillic”) and scripts (e.g. uk-Latn “Ukrainian transcribed to Latin”)
- Ability to create a hierarchy of concepts. This is used for hierarchical query expansion (see below).
- (Optional) Linguistic features of labels, e.g. gender
- Unify/concatenate gender-specific labels to have a unique preferred label per language (e.g. "Homosexueller Mann / Homosexuelle Frau"@de). Keep the gender-specific labels as alternate labels (e.g. "Homosexueller Mann"@de is Masculine vs "Homosexuelle Frau"@de is Feminine). See [Alexiev 2015c] sec.6.4 for analysis.
- (Optional) Related concepts (skos:related)
- Matching concepts in other thesauri (e.g. skos:exactMatch, skos:closeMatch). If EHRI will keep the current fragmentation into several thesauri, this is important; otherwise it is optional.

Main functional requirements

- Enable conceptual (semantic search):
  - Multilingual and synonym search: Search by one label of a concept (after picking the right concept using auto-complete) should find all documents indexed with any of its labels.
    - Hierarchical expansion (OPTIONAL): when searching for a concept (e.g. “Everyday camp life”), the user should find documents that are indexed with narrower terms as well (e.g. “Cooks”). This should be an option since the expansion of query results may not be desired by the researcher. For this to be useful, thesaurus hierarchies should be constructed properly (Broader-Than-Generic for concepts, and Broader-Than-Partitive for places, see [Alexiev 2015d])
- Provide auto-completion interface to be used for two purposes:
  - conceptual search
  - to enable manual cataloguing of archival descriptions by CHIs willing to use the EHRI concepts. (Note: using the LOD publication, such CHI could implement auto-completion themselves, but providing it at the EHRI portal will facilitate adoption).
- Provide a gazetteer (all labels) to be used for NER tasks (concept extraction)
- Publish or export as LOD in SKOS and SKOS-XL

Currently the concepts are fragmented in several thesauri:

- EHRI thesaurus
- Terezin keywords

- NIOD Trefwoorden

Keeping separate thesauri provides less value to the user, since when the same (or closely related) concepts appear in several thesauri, he/she would not have clear guidance as to which of them to select:

- for searching, or
- for manual cataloguing (adding an access point to a description).

### 7.1.1.1 Properties

- Labels in different languages
- Hierarchical relationships
- Associative relationships
- Scope notes (if possible)

### 7.1.2 Places

EHRI guidelines state that Geonames should be used as a place authority. Not only is Geonames a large place gazetteer (over 9M places), but it has a very useful place hierarchy (gn:parentFeature) that can be used for semantic search, e.g. a researcher looking for documents about a super-place (e.g. Poland) will likely be interested in documents indexed with sub-places thereof (e.g. Lodz).

Geonames coreferencing was not done in EHRI-1. We started the development of a rather sophisticated Geonames coreferencing service with the following features:

- Uses all place labels from Geonames
- Uses ancestor places to disambiguate (place name ambiguity is very common)
- Discards certain place types, e.g. Farms and Hotels
- Uses Population to rank candidates

We analyzed the coverage of this service on placeAccess access points, and the results are very encouraging. (We have not yet analyzed coverage on link types other than placeAccess, nor precision).

Out of 14,946 unique compound placeAccess, the Geonames reconciliation service recognizes 13,310 or 89%.

- e.g. the most popular recognized places are:
  - 13,870 <http://sws.geonames.org/3064268/> Terezín
  - 2,391 <http://sws.geonames.org/798544/> Poland
  - 2,274 <http://sws.geonames.org/3067696/> Prag
  - 2,150 <http://sws.geonames.org/2921044/> Germany
  - 1,984 <http://sws.geonames.org/2782113/> Austria
  - 1,420 <http://sws.geonames.org/2761369/> Wien,Vienna,Austria
  - 1,361 <http://sws.geonames.org/2750405/> Nederland
  - 1,310 <http://sws.geonames.org/5128581/> New York,New York,United States
  - 1,247 <http://sws.geonames.org/3067695/> Praha
  - 1,213 <http://sws.geonames.org/6252001/> United States,Emigration and immigration
  - 1,149 <http://sws.geonames.org/3079102/> Bohušovice nad Ohří
  - 1,040 <http://sws.geonames.org/3078610/> Brno

- (Note: Prag is recognized as geonames: 3067696 the city, while Praha is recognized as geonames: 3067695 is the first-order administrative division, but this is a minor imprecision).

These 13,310 place names are recognized as 5,567 places.

- e.g. both Poland and Polen are recognized as <http://sws.geonames.org/798544/>.
- However, such a large reduction from place names to Geonames places is suspicious. We are analyzing the reasons, and in many cases the reason is that cities are missed (only countries are recognized). e.g.:
  - 232 <http://sws.geonames.org/798544/> Lwow,Lwow,Lwow,Poland
  - 150 <http://sws.geonames.org/798544/> Vilna,Wilno,Wilno,Poland
  - 90 <http://sws.geonames.org/798544/> Wilno,Wilno,Wilno,Poland
- Perhaps the reason is historic change: today these cities are Lviv in Ukraine and Vilnius in Latvia. But the above names are (almost) unambiguous in Geonames, so they should be recognized despite the contradictory country (Poland).

1,636 placeAccess strings (11%) are not recognized in Geonames.

- Of them, 751 are “sub-city” places, such as Kaserne, Strasse, Gasse, Straat, Street, Park, Hotel or street in a particular quarter (e.g. Prag XII,Slezská 109). They do not appear in Geonames and should be left alone in the Terezin Places thesaurus.

The remaining 885 are a mixed bag that needs to be worked out:

- Historic place names such as British Mandate for Palestine, also written as Mandate Palestine or even **Mandatory** Palestine
- Unrecognized camps, e.g. Maly Trostenec, that should be recognized
- Places with a Nazi designation “Kreis” (e.g. Kreis Kiew) that should simply be ignored
- Sub-city features that are hard to recognize, e.g. Lyceum voor Joodse leerlingen aan de Stadstimmertuinen in Amsterdam; Room No,410 Room No,414
- Murder sites, such as Lopuchowo, forest. Unfortunately this is ambiguous: there are [4 places called Lopuchowo](#) (all are in Poland), and one needs to read about the [Tykocin pogrom](#) in Wikipedia to learn that the right one is in this hierarchy [Poland> Podlasie> Powiat białostocki> Gmina Tykocin> Łopuchowo](#)
- The murder site Lida Forest is even worse: there are some [40 places called Lida](#) (10 in China alone!) and from a cursory reading of Wikipedia we could not figure out where it is. This underscores the need to strengthen the information about Camps, Ghettos and Murder Sites (see section 7.1.3)
- Concepts that are place **types**, not specific places, e.g. Mass graves, Krematorium, Schleusenmühle (sluice mill)
- Concepts that have nothing to do with places, e.g. Mass murder, Oral history
- Codes such as Q711; Knabenheim,L 417; Kaffeehaus,Q 418
- Misspellings like Mass,Bostom (that must be Boston, Massachusetts) and Rotrterdam

We plan to enhance the service in several ways, including extraction of multiple places from compound access point (currently it extracts only 1 most likely place), increasing coverage, handling more special evaluating precision, etc.

## Difficulties

- Many common words (e.g. Temple, Yad...) appear as place names (many of them in the US). Currently we have a short blacklist of common words, but we are experimenting with word clusters extracted from Oral History interviews using machine learning (Neural Networks) through the word2vec program in WP14. For example, these clusters predict that the word Drama is more related to Music, Performance, or Theatre than to city names; so it is not very likely to indicate the Greek city of Drama.
- Geonames does not cover Terezin places like Streets, Kazernes (barracks), etc. This is easy to handle: we should just preserve the Terezin places authority. We will have 2 disconnected Place authorities, but that should not pose any significant problems. Geonames does not cover historic places well, e.g. Yugoslavia and Czechoslovakia appear, but there is no place hierarchy under them (the hierarchy is under the modern countries: Serbia, Montenegro, Czech Republic, Slovakia, etc.). Adding the historic hierarchy in Geonames would not be accepted by the Geonames community. The same probably holds of names like British Mandate for Palestine. So we should do such historic additions on a local RDF (semantic data) copy. This makes it necessary to use Ontotext GraphDB and presents a data maintenance task to enable synchronization with new Geonames versions.
- Geonames has some historic acronyms (e.g. CSSR for Czechoslovakia), but not others (e.g. both USSR and SSSR for Union of Soviet Socialist Republics are missing). We can add these directly to Geonames, which welcomes crowd-sourced collaboration.
- Geonames would be inadequate for “Nazi geography” places (changed borders and place names when an area was occupied). After the missing coverage has been analysed we plan to discuss alternative sources with experts in the area.
- It would be quite difficult to determine in all cases whether an access point is about a camp/ghetto or the associated place. So we propose to accept this commingling of meaning.

### 7.1.2.1 Geonames

Although there is not an authority for generic place names in EHRI, we have observed that indexers of CHIs need to add place access points to the documents. This leads to some errors, such as:

- National reports have the name of the respective country as titles. They have been used to add place access points to collections.
- Ghettos and camps have often the name of cities. We have cases in which a camp (Berlin in the example below) has been introduced as access point presumably meaning the city.

Here one can see a good sample of both mistakes: [https://portal.ehri-project.eu/units/lt-002880-f\\_1398](https://portal.ehri-project.eu/units/lt-002880-f_1398)

Geonames is a large LOD resource of geographic features, containing more than 9M places, ranging from inhabited places to rivers, mountains, oceans, etc.



EHRI-1 Standards have mandated that places should be coreferenced to Geonames. However, this guideline has not been implemented.

We should use Geonames as a valuable resource that can provide the backbone of all place information in EHRI. There is little point for EHRI to keep its own Place vocabulary, because that would duplicate unnecessarily all the great effort that went into creating Geonames.

- Coreferencing to Geonames can help deduplication and data cleaning, thus fixing the problem that ingested Access Points, including at least 23 ways of spelling Lodz.
- Geonames includes an extensive place hierarchy (the `gn:parentFeature` property) that can be used profitably for query expansion, e.g. "Find all archival descriptions that mention Poland or any place in Poland".

### 7.1.3 Camps, Ghettos, Murder Sites

- EHRI publishes relatively little data about Camps and Ghettos. For example, consider May Trostinec (<http://data.ehri-project.eu/ehri-camps.rdf#2030>): the only available data is merely the label:
  - `skos:prefLabel "Maly Trostinec"@de-latn`

[Wikipedia](#) publishes a lot more. References are provided for many of the facts, but the information is **not structured**:

- General
  - names: Maly Trostinets, Maly Trastsianiets, Trasciane, Малы Трасцянец, Maly Tras'tsyanyets, Малый Тростенец, Maly Trostinez, Maly Trostenez, Maly Trostinec, Klein Trostenez
  - location: outskirts of Minsk
  - admin district: Reichskommissariat Ostland
  - established: 10 May 1942
- Victims
  - victim countries: predominantly Belarus (inferred, not explicitly stated). Also Austria, Germany, Czech Republic
  - victim places: predominantly Minsk. Also Berlin, Hanover, Dortmund, Münster, Düsseldorf, Cologne, Frankfurt am Main, Kassel, Stuttgart, Nuremberg, Munich, Breslau, Königsberg, Vienna, Prague, Brünn, Theresienstadt
  - known victims:
    - Vincent Hadleŭski (Wincenty Godlewski): arrested in Minsk on December 24, 1942 and shot at Trascianiec the same day.
    - Norbert Jokl (debated)
    - Margarete Hilferding (in transit to the camp from Terezín)
    - Grete Forst
    - Cora Berliner (most likely)
- Perpetrators and Grounds
  - Murder sites (killing grounds): Blagovshchina (Благовщина) forest, Shashkovka (Шашковка) forest
  - Perpetrators (and their fate):
    - lead: SS Unterscharführer Heinrich Eiche (fled to Argentina after the war and all trace of him was lost)
    - Eduard Strauch (died in Belgian prison in 1955).

Rottenführer Otto Erich Drews (sentenced to life imprisonment by a court in Hamburg in 1968)  
Revieroberleutnant Otto Hugo Goldapp (In 1968 the Court in Hamburg sentenced to life imprisonment)  
Hauptsturmführer Max Hermann Richard Krahn (In 1968 the Court in Hamburg sentenced to life imprisonment)  
Heinrich Seetzen (committed suicide in a British POW camp)  
Gerhard Maywald (settled after the war in West Germany; On August 4, 1977 sentenced to 4 years imprisonment)  
Jewish Sonderkommando 1005

[Wikidata](#) provides the following **structured** information:

- Names and Wikipedia links in the following languages:  
Беларуская, Беларуская (тарашкевіца)Deutsch, Español, Français, ,Čeština, Dansk , עברית ,Frysk, Italiano, Nederlands, Norsk bokmål, Polski, Português, Русский, Српски / srpski, Suomi, Svenska, Українська, 中文
- additional aliases, e.g. Vernichtungslager Maly Trostinez, KZ Maly Trostinez, Blagowschtschina
- country: Belarus
- location: 53°51'3"N, 27°42'17"E
- Authority IDs: Geonames, VIAF, Freebase

[DBpedia](#) provides the following **structured** information:

- links to Wikidata, Geonames, Freebase, different Wikipedias
- coordinates
- a few more aliases:  
Maly\_Tras'tsyanyets  
Maly\_Tras'tsyanyets\_camp  
Maly\_Tras'tsyanyets\_concentration\_camp  
Maly\_Tras'tsyanyets\_extermination\_camp
- the fact that it is DeathPlace. This comes from the articles about these people (i.e. inverse links):  
dbr:Margarete\_Hilferding  
dbr:Grete\_Forst  
dbr:Vincent\_Hadleŭski
- categories:  
dbc:World\_War\_II\_sites\_of\_Nazi\_Germany  
dbc:Geography\_of\_Minsk  
dbc:History\_of\_Belarus\_(1939–1945)  
dbc:History\_of\_Minsk  
dbc:Maly\_Trostenets\_extermination\_camp  
dbc:The\_Holocaust\_in\_Belarus  
dbc:World\_War\_II\_sites\_in\_Belarus  
dbc:Belarus\_in\_World\_War\_II

Some of the information above can be controversial, disputed, unverified, etc., but the names/labels, geographic coordinates, establishment time, and possibly victim origin information (call them **factoids**) are probably true, and can be useful for various NLP tasks.

We believe that EHRI should develop much richer authoritative data about **camp**s and **ghettos**, and add **murder sites** (all of which appear often in access points). The data should be limited to factoids (non-controversial data). Some possible approaches could be:

- extracting factoids from Camp and Ghetto encyclopedias by EHRI partners and use Wikidata (see 9.2) as a semantic integration platform
- Holocaust researchers use Wikidata to curate and add more structured data. Wikidata is open for editing by anyone, but we do not believe there will be “editorial wars” or falsification of data if we limit the scope to factoids only.
- Try specialized NLP over articles about Camps and Ghettos from Wikipedia or EHRI partners to try to extract factoids

### 7.1.3.1 Properties

#### Properties of camps

- Labels in different languages
- Geodata
- Closest place (Geonames)
- Initial date
- Final date
- Hierarchy (camp-subcamps)
- Link to Wikipedia/Wikidata (then a lot of the above information can come from there)

#### Properties of ghettos

- Labels in different languages
- Geodata
- Closest place (Geonames)
- Initial date
- Final date
- Link to Wikipedia/Wikidata
- Link to Yad Vashem encyclopedia (?)

### 7.1.4 Person entities

Person entities are distributed into different datasets. The first set, created as EAC records contains 15 fields for person. This set has been incremented with other entities provided by WP15 during their identification work.

The other sets are authority sets produced for the research guides about Terezin and Jewish Councils, and a list of Terezin victims. They have a different degree of granularity in the descriptions.

Out of 36,736 personAccess access points used in archival descriptions, only 9,331 (25%) are created as an Authority object (historicalAgent). 13,952 personAccess access points (38%) have some sort of year recorded. Unlike Geonames, there is no reference dataset of people in the Holocaust domain to coreference against. We could explore coreferencing to the following sources:

- USHMM Persons: 3.2M person records plus 1M names of related people, including dates and places. A comprehensive overview is available on google docs in [USHMM Files](#) and statistics in [USHMM Persons](#). This dataset is not deduplicated, but WP13 will be working with it on the research case of Jewish Social Networks, so it will attempt deduplication. Since this research case will work on prosopographic information (integrating events, places, related people), USHMM Persons can be a very valuable resource to link to.
- VIAF: comprehensive information about 11M “notable” people (published or were described in some work). Includes basic life data, many labels (name forms) and bibliographies.
- Wikidata: about 2M records about notable people, of which half are new (not coreferenced to VIAF). Includes additional data and labels (name forms), see [Alexiev 2015b].

We recommend for EHRI to:

- Attempt such coreferencing, although the relatively small number of available dates will make it difficult
- EAC itself is not very “semantically oriented” because it records mostly names not links. As a minimal improvement, extend the EHRI EAC editor (Eddy) to:
  - Allow linking to external LOD sources
  - Once such a link is established, copy data from the external source to allow faster data population (similar to what the tools RAMP<sup>35</sup> and xEAC<sup>36</sup> are doing)
  - Allow typed relations between people, going beyond EAC (e.g. mother/father rather than just associative).
- WP14 and WP13 could try specialized NLP over EAD biographical information (<bioghist>) to extract dates, events and related people. Bioghist is, in many cases, the most comprehensive information that EHRI has about a person, but it is not structured.

Manual editing of 25-36k Person records will be very effort-intensive. Because of limited resources, we considered to focus on creators of archival materials (creatorAccess). Looking at the statistics in section 3.1.4.2, there are 8,805 unique creatorAccess, of which 2,593 (29.5%) are created as Authority objects (historicalAgent), 6,212 are mere strings (70.5%) and 576 appear as both. Of these, only 885 (10%) have some sort of year recorded. Many of them are corporateBodies not Persons, so this idea needs to be reconsidered. We have not yet analysed how many names appear as both personAccess and creatorAccess (but surely the coverage is not total, i.e. there are Persons that are only marked creatorAccess without having a second access point marked personAccess).

In summary, working with Agents (Persons and CorporateBodies) is one of the hardest

<sup>35</sup> <https://github.com/UMiamiLibraries/RAMP>

<sup>36</sup> <https://github.com/ewg118/xEAC>

Authority tasks.

#### **7.1.4.1 Properties**

- First name and last name
- Alternative forms of name
- Existence range (birth and death date)
- Birth and death place
- Other relevant dates (e.g. of arrest)
- Other relevant places (e.g. of arrest)
- Occupation and role
- Biographic history
- Associative relationships
- Bibliographic information:
  - Books or papers written by the person
  - Books or papers about the person

#### **7.1.5 Corporate bodies**

There are 5,142 access point with link type `corporateBodyAccess`, of which only 222 (4%) are created as authority objects (`historicalAgent`). However, it appears that more than half of the `creatorAccess` links described in the previous section are actually corporate bodies, so the number may rise to 9.5k, out of which 1.5k or 15.8% are authority objects. A lot of the corporate bodies are poorly described.

We could try the same approach of linking to external LOD and (semi)automatically enriching with LOD data as described in the previous section for Persons. But this will be more complicated, since there are many ways to spell the name of a corporate body.

##### **7.1.5.1 Properties**

- Name
- Alternative names
- Existence range
- Place
- Biographic history
- Associative relationships
- Bibliographic information
  - Books or papers written by the corporate body
  - Books or papers about the corporate body

#### **7.1.6 Events**

Import them to the portal after the current problems (there is not a template for events) have been solved.

##### **7.1.6.1.1 Properties**

- Label of the event
- Time

- Agent/Actors
- Place
- Associative relationships
- Hierarchical relationships

### **7.1.7 Administrative districts**

No modification will be introduced in this phase. EAC edition in the administrative interface will be enabled.

#### **7.1.7.1 Properties**

- Name of the administrative district
- Hierarchical links

## **7.2 Transformation and Curation of new introduced entries/instances**

### **7.2.1 Concepts**

- Validate proposed new thesaurus concepts
- Validate concepts extracted from EADs during ingestion process
- Maintain consistency in the hierarchy of concepts
- Provide translations
- Validate proposed translations
- Definition of the concepts/scope notes

### **7.2.2 Person entities**

- Validate person entities extracted from EAD access points
- Validate person entities extracted using NLP technologies
- Validate person entities provided by users (researchers and CHIs)
- Import them into the selected edition tool
- Enrich description of the person through:
  - Prosopography and text analytics
  - Automatic enrichment procedures (discuss the use of RAMP and xEAC)
  - Manual annotation and final validation
- Import into the portal
- Model an adequate edition interface using the actual editor
- Publication as LOD

### **7.2.3 Corporate bodies**

- Validate corporate body entities extracted from EAD access points
- Validate corporate body entities extracted using NLP
- Validate corporate body entities provided by users (researchers and CHIs)
- Import them into the selected edition tool
- Enrich description of the person through:

- Analytics. A collaboration with WP13 will be explored.
- Manual annotation and final validation
- Import into the portal
- Model an adequate edition interface using the current editor
- Publication as LOD

### 2.1.1 Events

The project has to decide which events are relevant.

- Model suitable edition interface
- Links to places, agents

## 7.3 Creation and modelling of new vocabularies

EHRI will need to create extra thesauri for the description of the archival material, collections and relationships of the portal. Since most of them are already standardized, we expect to obtain them from various external sources. Examples of those vocabularies are:

- Type of material (or genre). We can import a lot from the LCSH Genre thesaurus
- Type of link between archival collections (e.g. original-copy)

Other small thesauri can be created to allow systematic editing of Person properties. For example, we have seen these Authority lists in USHMM Person data:

- Gender
- Marital status (married, widowed, engaged, ...)
- Ethnicity
- Citizenship/nationality
- Political ideology/orientation/identity
- Religion
- Holocaust fate (survived, murdered, buried...)

A final selection of small vocabularies will be made after the final decision on the properties associated with Persons and the necessities of the material currently imported into the portal.



## 8 Publishing the Authorities

We foresee the publication of authorities in several formats:

- Linked Open Data for automatic consumption by other projects that want to link to the EHRI authorities
- EAC CPF (XML) for archive-oriented agent data
- Interactive tools for easier use by users

Other formats could also be considered, such as TBX, if there is a demand for them.

### 8.1 EAC CPF Formats

Agent (Person and Organization) data should be published in EAC CPF XML for consumption by archives. EHRI-1 has developed an export from the EHRI database (neo4j) that needs to be validated and perhaps extended<sup>37</sup>.

### 8.2 Linked Open Data

The main shortcoming of EAD and EAC CPF is that they are not very semantic:

- EAD access points are strings, not links to global LOD entities
- CPF does not provide links to people but only their names
- Events are local to each CPF record, and do not refer to global data about events; e.g. it is impossible to say that two people participated in the same event by using global Person URLs.

We recommend that EHRI publishes a lot of the Authority data in semantic format as LOD.

- Unlike the current approach that serves large files, we should serve only the triples for the individual Authority entity requested
- We should serve multiple formats (HTML plus multiple machine readable formats: RDF/XML, Turtle, JSON-LD, NTriples) using content negotiation. This means that ideally, entity URLs should have the form e.g. <http://ehri-project.eu/thesaurus/ehri/1234>, and redirect either to <http://portal.ehri-project.eu/thesaurus/ehri/1234> (if HTML is requested) or <http://data.ehri-project.eu/thesaurus/ehri/1234> (if RDF is requested)
- We should use permanent URLs (to be specified by T11.3) and be properly published as per LOD best practices

Semantic data should be published using appropriate ontologies:

- For concepts: SKOS + SKOSXL
- For historic place names and hierarchy: the Geonames ontology
- For people and corporate bodies: FOAF/Bio Schema
- For events, dated, related people: the EAC CPF ontology as proposed by [Mazzini, Ricci 2011] and [Eito-Brun, 2014], or CIDOC CRM

---

<sup>37</sup> That has a low priority, since it is not very likely that any archives would consume EHRI EAC.

### 8.3 Auto-Completion

EHRI should provide auto-completion interfaces to:

- Allow the portal user to do semantic search. Both vetted (official) and candidate thesauri should be consulted (since existing descriptions use access points that will not make it into the vetted thesaurus, at least not in the short-term).
- Allow the archivist to use EHRI thesaurus entries interactively during manual cataloguing of new descriptions. Only vetted thesauri should be consulted for this.

These interfaces are non-trivial since:

- They need to consult multiple authorities, e.g. for places, both Geonames and internal thesauri like Terezin Places.
- They should display enough information to let the user understand the entry (e.g. show it in context), yet little enough to allow user-friendly display of the auto-complete list.
- They need to implement full-text-Search that works across multiple languages
- They need to be highly performant to provide a user-friendly experience.

## 9 Tools and Applications

### 9.1 Edition on the administrative interface of the portal

The EHRI portal offers an interface for the edition of entries of some categories of the vocabularies. Although the interface has to be improved in order to be able to handle the manual edition of EAC vocabularies, it has been already extensively used by EHRI partners to add new personalities and corporate bodies to the portal.

We recommend to extend the actual interface to handle the following entity types:

- **Person:** Extend the actual model to provide a richer set of relationships for interlinking. Here the editor must continue to be EAC compliant.
- **Corporate bodies:** follow the EAC standard
- **Administrative districts:** For the moment these are just strings linked by hierarchical relationships. The list seems to be already closed, but it is not easy to foresee which kind of properties will be needed. Since administrative districts have an institutional character, we recommend to model them using the same template as corporate bodies, and allow to link them to the same entities in the domain.
- **Events:** The interface will need an extra template for the edition of events with fields compliant with the SEM ontology used to model the events in EHRI-1.

### 9.2 Wikidata as a Semantic Integration and Editing Platform

Wikidata is a global knowledge base of “everything” and one of the most active projects of the Wikimedia Foundation. It has information on about 16M entities, covering all 280 Wikipedia language editions (5M come from en.wiki), plus additional entities. It has a comprehensive data model where one can add not only entity attributes (data values) and relations (to other entities), but also statement references and qualifiers.

For example, when adding the population of a city, it would be appropriate to record:

- reference (e.g. website showing the number)
- originating agent (e.g. national statistical agency)
- method of obtaining (census or estimation)
- area covered (metro or greater city area), etc.

#### 9.2.1 Wikidata for Editing

Wikidata is easy to use as an editing tool, [see Alexiev 2015]. It has already been used by many GLAM institutions, e.g.:

- The WikiProject Sum of All Paintings has the ambition to build a catalogue raisonné of all paintings in the world - see [Wikidata 2015a] for project pages and [Wikidata 2015a] for a description of benefits to museums.
- A group of Flemish museums and art collections publish and edit metadata of their art collections (estimated: 35,000 artworks in total) on Wikidata [PACKED 2015a], [PACKED 2015b].

Wikidata could be used in two deployment modes:

- The central installation at <http://www.wikidata.org>. This is easiest to use since it is under constant development, and has a great community of users that can provide assistance, and developers that can create data import/integration tools (**bots**).

- A custom installation, e.g. as used by the EAGLE project [Orlandi et al 2014, chapters 1 and 12]. EAGLE deals with epigraphic data (stone inscriptions). Because these are relatively simple objects (no “nested parts”), they used the Wikidata software (Wikibase) for semantic data integration from many providers, and for editing. EAGLE is the second project (apart from Wikidata) to use the framework. The technical work was done by Wikimedia Italy.

Using the central installation has some consequences:

- Anyone can edit the data, and there may be doubts about its authority or veracity. This may seem daunting at first, but distributed editing has worked surprisingly well, as 10 years of experience with Wikipedia clearly show. Wikidata places a special emphasis on referencing statements to primary sources, with special efforts for tooling and referencing campaigns.
- Adding very many entries (e.g. several million person records) will perhaps meet with opposition from the community. But Wikidata currently does not enforce notability guidelines strongly (which is just one of the reasons why GLAMs can work a lot more easily with Wikidata than Wikipedia).
- Data on victims has privacy implications (subject to the legal agreements between EHRI partners), so it should not be exposed openly.

We would recommend that EHRI consider using the central Wikidata installation for Camps and Ghettos only (as an initial experiment), see below.

## 9.2.2 Wikidata for Integration and Coreferencing

Even more importantly, Wikidata can be used as an integration platform. It has a flexible data model that can be extended for any purpose, can easily ingest data in various formats, and can correlate the data. Wikidata has proven an effective data integration platform to the EAGLE project (see above), which used it for integrating data from many partner institutions.

Wikidata [Mix-n-Match](#) [Manske 2014] is an excellent and widely used tool for coreferencing various authority lists to Wikidata. It has great promise in realizing the Holy Grail of librarians and authority control specialists: a world-wide integrated authority file.

- Over 100 authority databases are loaded for coreferencing, and several million entities are coreferenced.
- See some news and screen shots at <https://twitter.com/hashtag/coreferencing>
- See the WikiProject Authority Control [Ontotext 2015] for more information and coordination activities.

When it comes to global person authorities in particular, Wikidata and VIAF are the only two that need to be considered: they dominate the “open data tradition” and the “library tradition” datasets respectively [Alexiev 2015b].

- Wikidata has information on about 3M people. VIAF has 33M records. Of them 1.8M are common with Wikidata.
- VIAF is now actively sourcing Wikidata, so the gap will be closed quickly.
- Wikidata provides easy export of coreference information through the [Beacon tool](#). This allows easy integration of authorities that are not integrated in VIAF, e.g. RKD Artists

## 9.2.3 Using Wikidata for Camps and Ghettos

Section 7.1.3 shows that public data about camps is much richer than what is available from EHRI. We think that it would be appropriate and easy for EHRI to use Wikidata to edit information about Camps and Ghettos.

As an experiment, we entered some of the information about Maly Trostenets described in section 7.1.3 and edited it in Wikidata (<https://www.wikidata.org/wiki/Q316109>). We were able to add about 10 labels and 15 facts within a few minutes. We did not try to add the following information, because it needs new properties to be proposed and agreed by the Wikidata editorial community:

- Victims typical place of origin (could twist property “journey origin” and use it for this)
- Killing grounds
- Perpetrators and their positions

The best way to see the information is in [Reasonator](#)

- The large number of labels can be very useful for NLP tasks
- Geographic coordinates and Geonames binding (id, administrative region Minsk, country Belarus) can be useful for geo exploration and reconstructing life histories

### Maly Trostenets (Q316109)

Концлагерь Мальный Тростенец | Тростенец | Малы Трасцянец | Канцлагер Трасцянец | Вёска Малы Трасцянец | Мемарыяльны парк Трасцянец | Логор смрти Мали Тростанец | Логор смрти Мали Тростенец | Maly Trostinek | Trostenez | KZ Maly Trostinez | Maly Trostinec | Maly Trostenez | Blagowschtschina | Trosteneц | Klein Trostenez | Vernichtungslager Maly Trostinez | Maly-Trostenets | Трасцянец | Maly Trostinets | Maly Trastjanets | Trasciane | Maly Trostinez | Maly Tras'tsyanyets | Malyy Trostenets

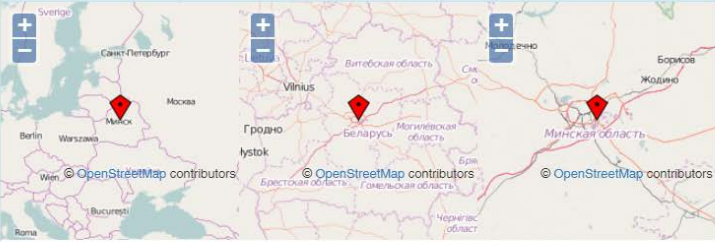
extermination camp near village in Belarus

**Location properties**

**instance** Village

**of** Extermination camp

**Maps**



Other Wikidata items within 15km | Geohack | TagInfo | Overpass | 53.85111111111111 / 27.70472222222222

**Location**

Name	Description
Belarus	country in Europe
Minsk	capital city of Belarus
<b>Maly Trostenets</b>	extermination camp near village in Belarus

Free images Google search

**External sources**

Freebase /m/02h7p7

GeoNames 625431

VIAF 236124295

**External identifiers**

**Wikimedia projects**

**Current language Wikipedias**

en Maly Trostenets extermination camp

**Big Wikipedias**

cs Malý Trostinec

da Maly Trostenets udryddelseslej

de Vernichtungslager Maly Trostinez

es Maly Trostenets

fi Maly Trastjanets

fr Maly Trostenets

it Campo di sterminio di Maly Trostenets

nl Maly Trostenets

pl Niemiecki obóz zagłady w Małym Trościeńcu

pt Maly Trostenets

ru Мальный Тростенец

sv Maly Trostenets

zh 玛丽特罗斯特内兹灭绝营

**Other Wikipedias**

be Трасцянец

be\_x\_old Трасцянец

fy Maly Trostenets

he ירידה לטבח "מא

no Maly Trostenets

oc Camp de Maly Trostenets

sr Логор Мали Тростанец

uk Малий Тростенець

- Links to related entities (in this case, people who died there) are also shown

D.11.2 Road Map Domain Vocabularies

Page 59

Other properties	
<b>heritage status</b>	cultural property of Belarus
<b>inception</b>	1942-05-10
<b>topic's main category</b>	Category:Maly Trostenets extermination camp Wikimedia category



From related items	
<b>place of death</b>	Cora Berliner German economist
	Dora Bromberger German painter
	Elsa Bienenfeld Austrian music historian
	Jeanette Schocken deutsche Kaufhausbesitzerin
	Oswald Levett Austrian author
	Vincent Hadleŭski Belarusian priest, publicist and politician
	Albert Katzenellenbogen deutscher Justizrat, Bankier und Opfer des Holocaust
	Alma Johanna Koenig Austrian translator and writer
	Grete Forst Soprano
	Margarete Hilferding Austrian MD & psychoanalyst

It is important to consistently add the statement “instance of: extermination camp” (or another agreed type) so it will be easy to get the data back after editing it in Wikidata.

### 9.3 VocBench

The EHRI portal has a section for editing controlled vocabularies, but it has its limitations: it cannot see the complete hierarchy, cannot move concepts in the hierarchy, cannot create links to external thesauri (e.g. skos:exactMatch, skos:closeMatch). Furthermore, the EHRI Thesaurus currently has serious quality problems, see section 3.1.5 (EHRI Thesaurus Quality).

Therefore EHRI should consider the deployment of a proper thesaurus management system for its Concept (Topic) thesauri.

[VocBench](#) is an open source thesaurus management system that works over SKOS/SKOSXL (the RDF ontology for representing thesauri) stored in Ontotext GraphDB as repository.

VocBench is a Terminology Management system developed by the United Nations Food and Agriculture Organization (FAO) and University of Roma Tor Vergata over a number of years, and is used extensively in production systems.

- Design work on VocBench began in 2004
- A first version was released in 2007
- A major rewrite was undertaken in 2011-2012. This version (VocBench 2) underwent beta testing in February 2013 and went into production in summer 2013. This version represents a major improvement in both the overall design and the flexibility of VocBench.

#### 9.3.1 VocBench Users

VocBench has supported a number of thesaurus management efforts at FAO:

- The AGROVOC thesaurus of 40k terms, with labels in up to 22 languages and 4 more languages in development
- The Biotechnology Glossary
- Land and Water



- FAO Topics
- Bibliographic metadata used in FAO

Some of these diverse datasets have required extensive customization, including additional fields. VocBench is one of the most important pillars of FAO's Linked Open Data strategy. VocBench has been released as open source and FAO encourages deployments by other institutions, and solicits contributions from third-party developers. Such collaborations may advance the development of VocBench further, to the benefit of all concerned parties.

Other relevant organizations interested or already using VocBench for the maintenance of their thesauri include:

- EU Documentation Office > EUROVOC
- EC Parliament Library
- European Environment Agency (EEA) > GEMET
- Scottish Government > Gov metadata
- Italian Senate > TESEO
- Harvard University > Unified Astronomy Thesaurus (UAT)
- Agence Nationale de la Recherche > Infrastructure nationale AnaEE France
- CABI
- United Nations Convention to Combat Desertification (UNCCD)
- Columbia University > IEDA Thesaurus

### 9.3.2 VocBench Features

- Native support for SKOS, the W3C standard for thesaurus representation
- Live SPARQL endpoints over thesauri and Linked Open Data integration, without having to export to SKOS
- Support for multiple repositories (triple-stores)
- Native support for OWLIM<sup>38</sup>. FAO has worked closely with Ontotext to fine-tune performance and resolve any outstanding issues.
- Modern and responsive user interface based on GWT
- OSGi-based architecture that allows flexibility and easy addition of modules
- A Java based architecture that allows easy translation of the user interface to any language
- Convenient access to semantic data (RDF, RDFS, and OWL) using the light-weight OWL Art API of University of Roma Tor Vergata. This component also includes semantic annotation and ontology enrichment (called Semantic Turkey).
- Full multilingual support of thesaurus data including all Unicode charsets
- Support for concurrent, distributed editing
- Editorial workflow supporting user roles and concept statuses (e.g. proposed, approved, published)
- Editorial rights by language
- Tracking of editorial changes and authorship of changes
- Activity reporting in the VocBench Workbench

In the appendix we provide illustrations of some VocBench features.

---

<sup>38</sup> Previous version of the GraphDB repository



## 9.4 SKOS Visualization (SKOSPlay)

There are various pieces of software that can visualize, browse or publish SKOS thesauri. One of the best ones is [SKOS Play](#). It is a free application to render and visualize thesauri, taxonomies or controlled vocabularies expressed in SKOS. With SKOS Play you can print Knowledge Organization Systems that use the SKOS data model in HTML or PDF documents, and visualize them in various graphical representations.

- Generate printable versions of thesaurus or knowledge organization systems
- Bridge the gap between SKOS data and data visualization provided by d3.js<sup>39</sup>
- Demonstrate and illustrate how some of the technologies of the web of data work
- Verify a vocabulary when working on it, validating it with domain experts, publishing it on the web

You can [try out a demo](#). It allows upload of own files (but only up to 5k concepts).

In the appendix we provide some examples from one of the EUROVOC microthesauri (conceptSchemes) and illustrations of some SKOSPlay features.

---

<sup>39</sup> JavaScript library for producing dynamic, interactive data visualizations in web browsers.

## 10 Effort Estimation and Time Table

Table 6 below gives an overview of all the tasks necessary to implement the work discussed in this roadmap, including an estimation on how many Person Months (PM) each task will take. Tasks have been prioritised according to the priorities suggested by the PMB.

We plan to deliver the results attending to the following milestones:

- Milestone M1 "Data cleaning and normalization": Oct 2016
- Milestone M2 "Curation tools": April 2017
- Milestone M3 "Integration of data and services": April 2018
- Milestone M4 "LOD and exports": January 2019

It is our expectation that we have enough resources to complete all tasks with priorities 1 and 2; there is a good chance that tasks with priority 3 will at least partially be implemented in the framework of EHRI-2; tasks with priority 4, finally, will be undertaken if resources are still available toward the end of the project.

WBS	Category	Task	Priority	Milestone	Responsible	Estimate (PM)
1.1	acc.points	Normalize, deduplicate (cluster)	P1	M1	ONTO, YV	0.5
1.2	acc.points	Decompose, discover type (eg "placeAccess" applies only to first atom)	P1	M1	ONTO, YV	1
1.3	acc.points	Coreference to internal thesauri (EHRI, Terezin, Terezin places...)	P1	M2	ONTO, YV	0.83
1.4	acc.points	Add non-coreferenced cluster as new Candidate Concept; add new term	P1	M3	ONTO, YV, KCL (for neo4j lookups)	1
1.5	acc.points	Coreference string placeAccess to Geonames or Terezin Places	P1	M2	ONTO	2
1.6	acc.points	Re-apply access points as Authorities to existing EADs	P1	M3	ONTO, YV	0.33
1.7	acc.points	Web service to handle acc.points of new EAD as part of Ingestion	P1	M2	ONTO (WP10.3 and WP11.4)	0.83
2.1	old thes	Deduplicate access points made as objects	P1	M2	YV, ONTO	0.38
2.2	old thes	Coreference placeAccess points made as objects to Geonames	P1	M1	ONTO	0.3
2.3	old thes	Update ALL EHRI data to new Authority URLs according to Permanent URL scheme	P1	M3	KCL, ONTO	0.75
2.4	old thes	Old Terms: correction of SKOS file using provenance information	P1	M1	YV	0.5
2.5	old thes	Camps and Ghettos: coreference to Wikidata, and add data to it	P3	M2	YV, ONTO	0.5
2.6	old thes	Camps and Ghettos: Plan and implement synchronization with the portal	P3	M3	KCL, ONTO	0.38
2.7	old thes	FAST keywords: classify into categories, coreference places to Geonames	P1	M1	YV	0.42
2.8	old thes	Import events into the portal	P1	M2	YV, KCL	0.08
2.9	old thes	Create suitable templates in the portal for presentation of entries	P2	M2	KCL	0.27
3.1	new thes	Load Geonames to RDF repo, manage local additions and data updates	P1	M1	ONTO	0.48
3.2	new thes	Create archival-related small vocabularies (genre, form, type of link..)	P3	M3	YV	0.5
3.3	new thes	Create content-related small vocabularies (nationality, religion, political affiliation)	P3	M3	YV, ONTO	0.5
4.1	agents	Terezin/Jewish Council persons, merge with EHRI persons	P1	M1	YV	0.5
4.2	agents	Terezin/Jewish Council corporate bodies, merge with EHRI corporate bodies	P1	M1	YV	0.5
4.3	agents	Coreference & enrich persons/corporate bodies with VIAF data	P2	M3	YV, ONTO	1.17
4.4	agents	Coreference personalities with USHMM Persons database	P2	M3	ONTO (WP13 and WP11)	1

4.5	agents	Extract person facts from <bioghist>	P4	M4	ONTO, YV, WP13/14 (research)	2.17
5.1	thesaurus	Deploy VocBench, add small customizations	P2	M1	ONTO	1.08
5.2	thesaurus	Reconfigure administrative interface	P1	M1	KCL	0.78
5.3	thesaurus	Semantic (conceptual) search	P2	M3	ONTO (WP11), KCL (WP7)	0.5
5.4	thesaurus	Auto-complete (for search & new cataloging): GraphDB (subjects), neo4j (agents), Geonames (places)	P2	M3	ONTO (WP11), KCL (WP7)	0.52
6.1	lod	Create RDF/LOD repository (GraphDB)	P4	M1	ONTO	0.37
6.2	lod	Publish vocabularies as LOD from GraphDB	P4	M4	ONTO	0.42
6.3	lod	RDF export of person/corporate bodies from neo4j	P3	M4	KCL, YV	0.4
7.1	pm	Project management, telcos, meetings	P1	ong	YV, ONTO	1.5
8.1	ed.board	Edition of vocabularies, validation of entries, training for members	P2	ong	YV, CDEC, ONTO	2
					<b>Total</b>	<b>24.45</b>
					Total Prio 1-2	19.22
					Total Prio 3	2.28
					Total Prio 4	2.96

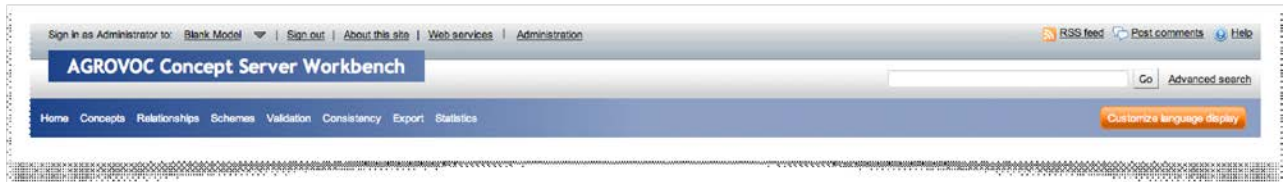
*Table 6: Tasks and required effort*



## Appendix

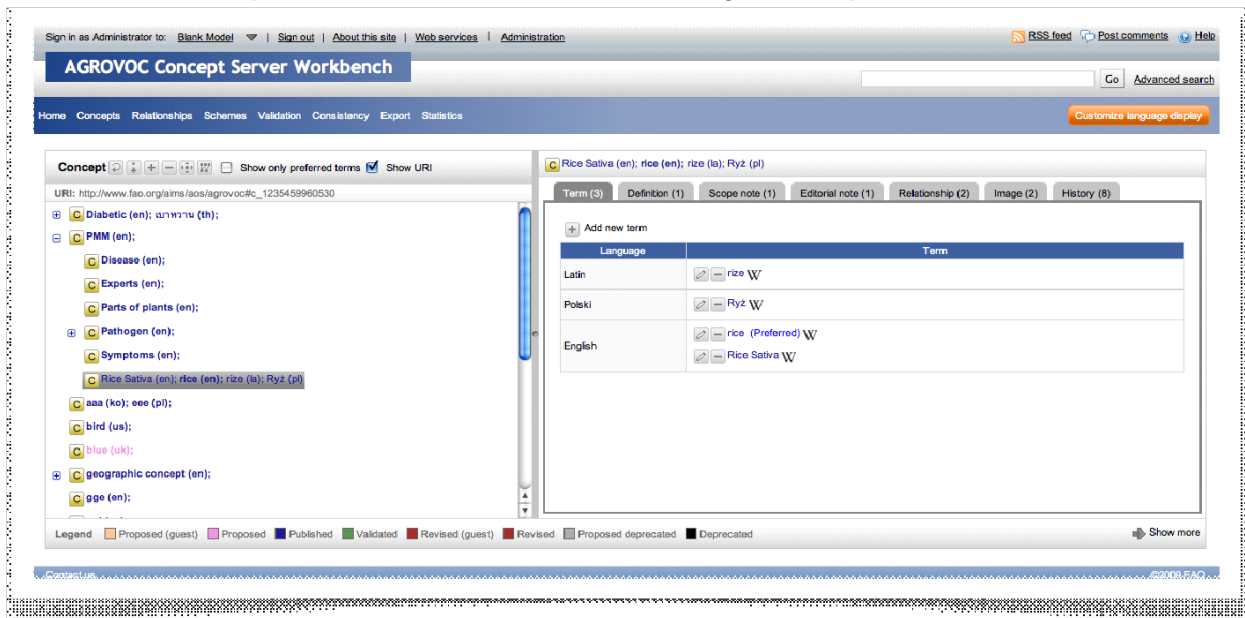
### Vocbench Features

#### Menu Bar



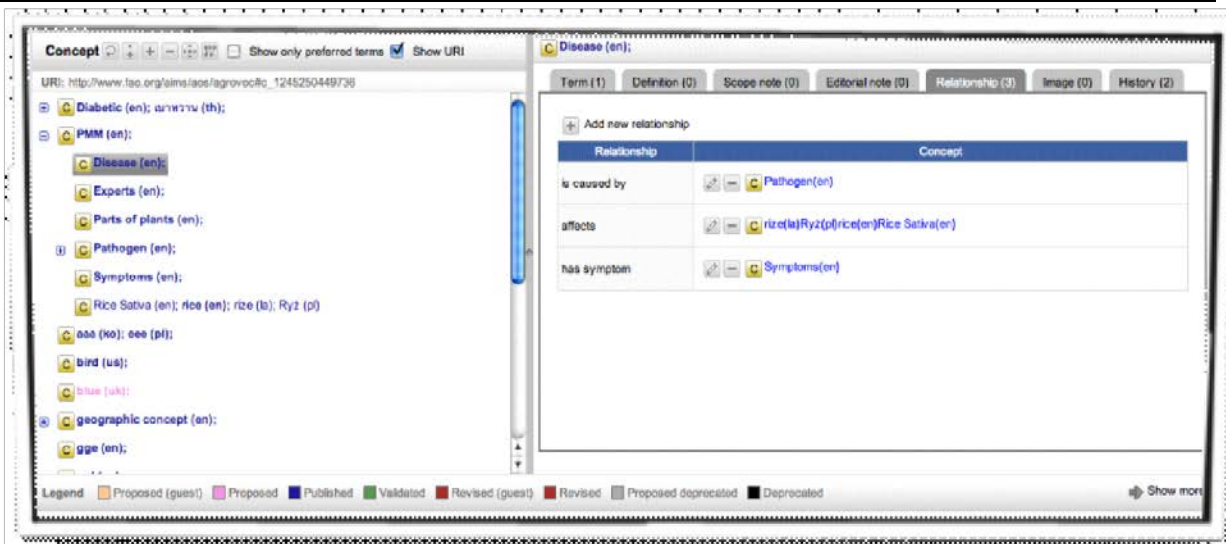
#### Concept Management

Notice links to Wikipedia and the various tabs describing a concept

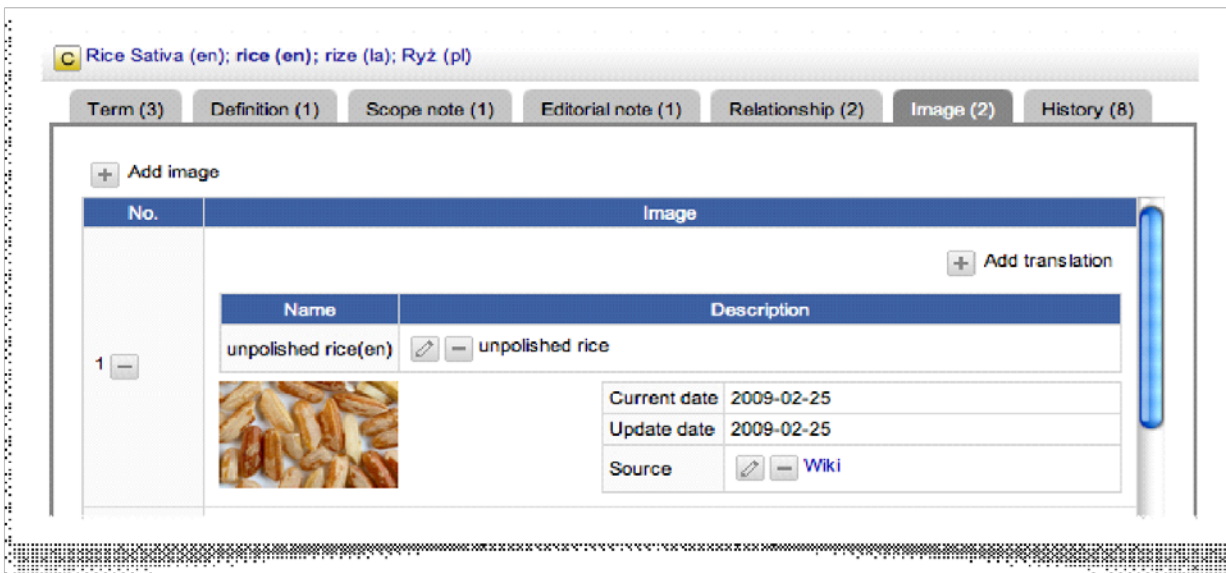


#### Concept Relations

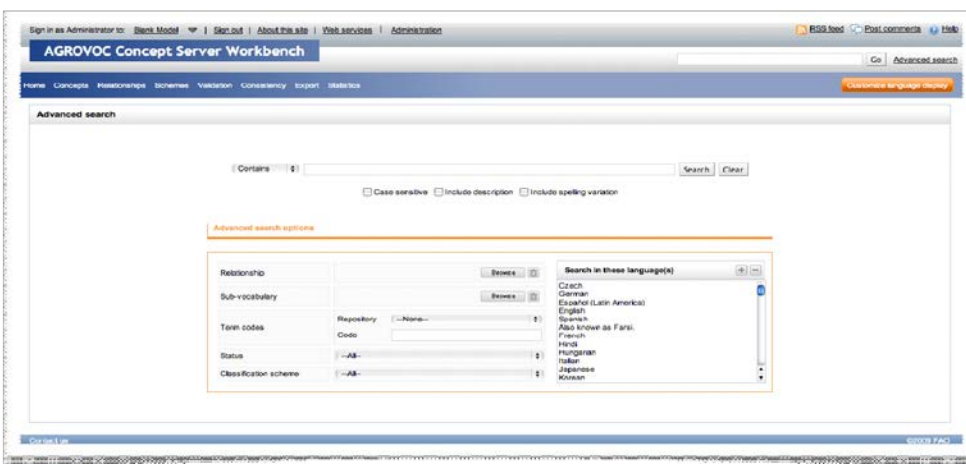
If defined, Inverse relations are created automatically



### Concept Image



### Search



### Recent Changes



Sign in as Administrator to: [Blank Model](#) | [Sign out](#) | [About this site](#) | [Web services](#) | [Administration](#) [RSS feed](#) [Post comments](#) [Help](#)

## AGROVOC Concept Server Workbench

[Advanced search](#)

Home | [Concepts](#) | [Relationships](#) | [Schemes](#) | [Validation](#) | [Consistency](#) | [Export](#) | [Statistics](#) [Customize language display](#)

---

[About AGROVOC](#) | [Glossary](#) | [Partners](#)

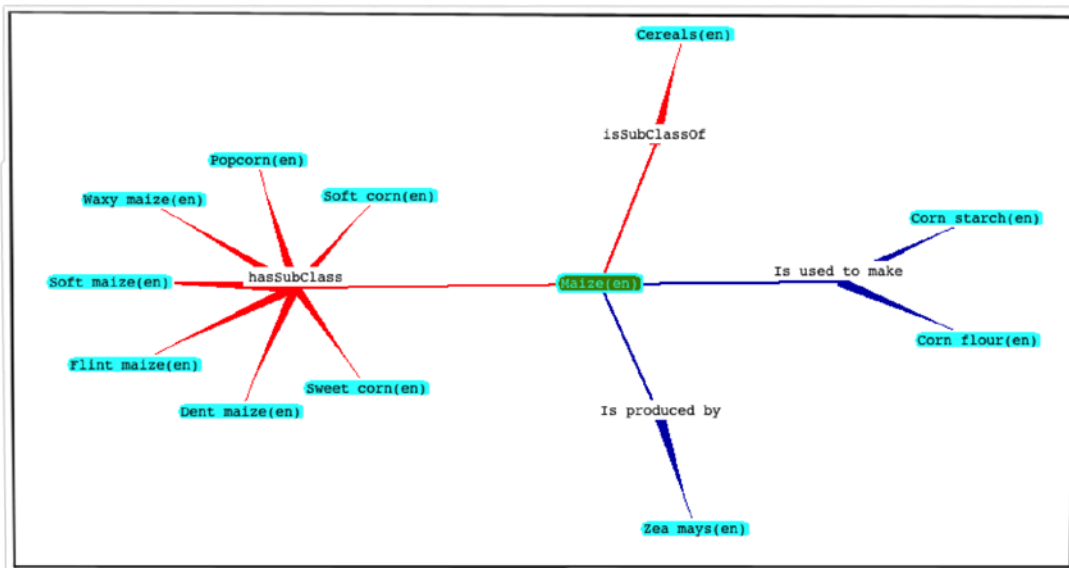
### Recent changes

Concept / Term / Relationship / Scheme	Change	Old value	Action	User	Date
Imma Subirats (subirats)			user-create	Guest	18-06-2009 19:01:50
<a href="#">blue (uk)</a>	<a href="#">asdf (en)</a>		concept-edi-definition-create	Administrator	18-06-2009 18:04:41
Top level concept (en)	<a href="#">blue (uk)</a>		concept-create	Administrator	18-06-2009 18:04:33
<a href="#">red (en)</a>	<a href="#">maroon (en)</a>		concept-create	Administrator	18-06-2009 09:47:32
<a href="#">red (en)</a>	<a href="#">maroon (en)</a>		validation-rejected - concept-cre	Administrator	18-06-2009 09:42:50
<a href="#">rice (la),Ryz (pl),rice (en),Rice Sadva (en)</a>	<a href="#">tisce (en)</a>		term-create	Administrator	17-06-2009 22:32:04
<a href="#">rice (la),rice (en),Ryz (pl),Rice Sadva (en)</a>	<a href="#">tisce (en)</a>		validation-rejected - term-create	Administrator	17-06-2009 22:27:31
<a href="#">ggg (sk)</a>	<a href="#">a11 (en)</a>		validation-accepted - concept-cr	Administrator	17-06-2009 22:25:46
<a href="#">ggg (sk)</a>	<a href="#">a12 (en)</a>		validation-accepted - concept-cr	Administrator	17-06-2009 22:25:46
<a href="#">ggg (sk)</a>	<a href="#">aa11 (en)</a>		validation-accepted - concept-cr	Administrator	17-06-2009 22:25:46
<a href="#">a11 (en)</a>		<a href="#">a11 (en)</a>	validation-accepted - concept-de	Administrator	17-06-2009 22:25:34
<a href="#">a11 (en)</a>		<a href="#">a11 (en)</a>	validation-accepted - concept-de	Administrator	17-06-2009 22:25:34
<a href="#">a12 (en)</a>		<a href="#">a12 (en)</a>	validation-accepted - concept-de	Administrator	17-06-2009 22:25:34
<a href="#">a999 (en)</a>	<a href="#">a911 (en)</a>		validation-accepted - concept-cr	Administrator	17-06-2009 22:25:34
<a href="#">testing (tr)</a>		<a href="#">testing (tr)</a>	validation-accepted - concept-de	Administrator	17-06-2009 22:25:26

1 of 25

Contact us ©2009 FAO

### Concept Graph View



### User help

العربية 中文 english français español

**Agricultural Information Management Standards (AIMS)**

**Help**

**WORKBENCH HELP INDEX**

**Glossary**

**Overview, using the home page and getting started**

- Basic functionality overview
- Overview of navigation and Workbench sections
- Setting languages for Workbench data (Concepts and classifications)
- Legend and icon overview
- Viewing recent changes in the Workbench (includes information about RSS feeds)
- Contribute your comments

**User roles and privileges**

- Overview of process
- Guests (Non-logged in users)
- Term editors (Terminologists)
- Ontology editors
- Validators
- Publishers
- Administrators

**Searching**

- Simple search
- Advanced search

**Concepts module**

**General:**

Food and Agriculture Organization of the United Nations  
for a world without hunger

Google Custom Search

**FAO Home**

**AIMS Home**

**News**

**Events**

**AOS : Registries**

- AGROVOC Concept Server
- AgMes Application Profiles and Ontologies
- Domain Ontologies
- Knowledge Organization Systems (KOS)
- KOS and Ontology Mappings

**Projects**

**Publications**

**Tools**

“ Interoperability, Reusability, and Cooperation ”

**SHORTCUTS**

- Registry of KOS
- Registry of metadata sets
- Registry of tools

**RELATED LINKS**

- AGRIS search engine
- AgriFeeds

**PARTNERSHIPS**

- CIARD

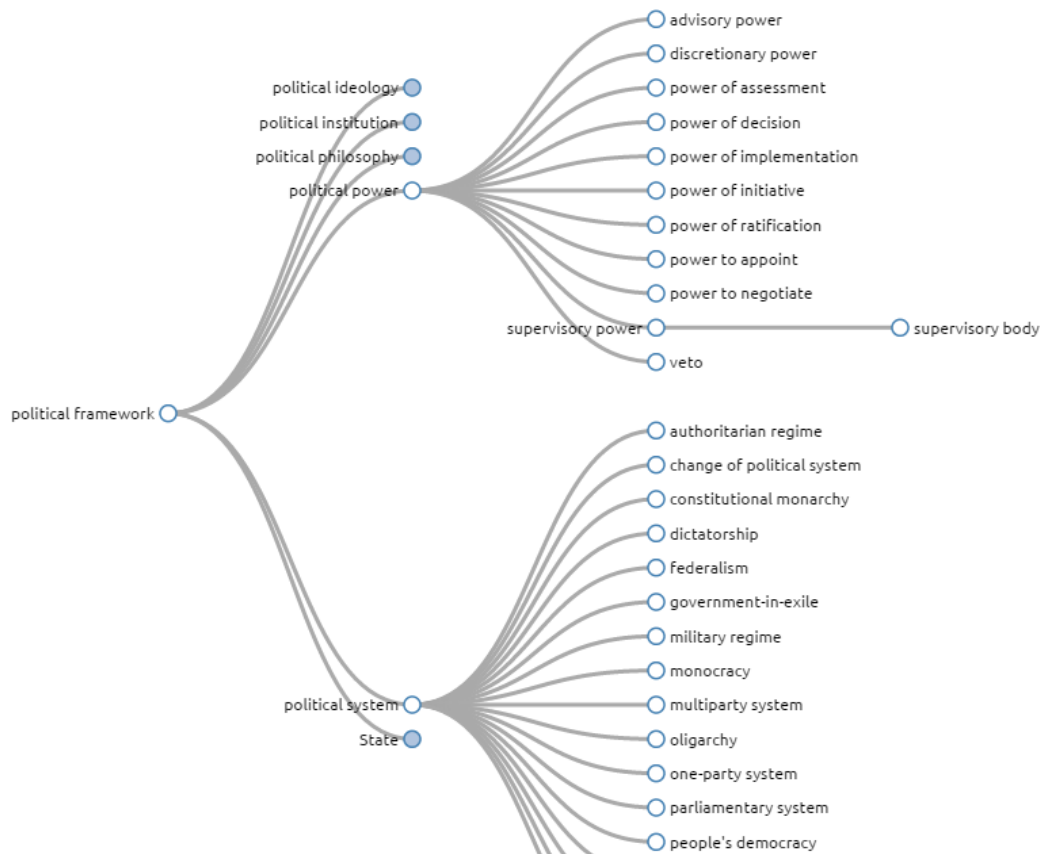
## SKOSplay examples

- Thesaurus Display

# 0406 political framework

- *absolute monarchy*
  - USE : **monocracy**
- **advisory power**
  - TT : **political power**
  - BT : **political power**
  - RT : **powers of parliament**
  - RT : **European Parliament**
- **anarchism**
  - TT : **political ideology**
  - BT : **political ideology**
- **authoritarian regime**
  - UF : **totalitarianism**
  - UF : **totalitarian regime**
  - TT : **political system**
  - BT : **political system**
- *autocracy*
  - USE : **monocracy**
- *Bundesland*
  - USE : **Federation State**
- *centralised State*
  - USE : **unitarian State**
- **change of political system**
  - TT : **political system**
  - BT : **political system**
  - RT : **political reform**
  - RT : **transition economy**

- Dendrogram



● Sunburst Diagram



## Glossary

- BIBO: Bibliographic Ontology <http://bibliontology.com>
- Bio: Vocabulary for Biographic Information <http://vocab.org/bio>
- CCS: CENDARI collection schema
- CHI: Collection Holding Institution
- CIDOC-CRM: CIDOC Conceptual Reference Model <http://www.cidoc-crm.org>
- DC: Dublin Core
- DCT: Dublin Core Terms
- EAC-CPF: Encoded Archival Context – Corporate Bodies, Persons, Families <http://eac.staatsbibliothek-berlin.de>
- EAD: Encoded Archival Description <https://www.loc.gov/ead>
- EAG: Encoded Archival guide for Holding Institutions
- FAST: Faceted Application of Subject Terminology <http://www.oclc.org/research/themes/data-science/fast.html>
- FOAF: Friend Of A Friend, ontology for the description of persons <http://www.foaf-project.org>
- Geonames: Geographical database <http://www.geonames.org>
- LCSH: Library of Congress Subject Headings <http://id.loc.gov/authorities/subjects.html>
- LOD: Linked Open Data
- MADS: Metadata Authority Description Schema <http://www.loc.gov/standards/mads>
- MODS: Metadata Object Description Schema <http://www.loc.gov/standards/mods>
- NLP: Natural Language Processing
- OSGi Architecture: Architecture proposed by the Open Service Gateway Initiative <https://www.osgi.org/developer/architecture/>
- OWL: Web Ontology Language <https://www.w3.org/2001/sw/wiki/OWL>
- PROV: vocabulary for provenance information <https://www.w3.org/TR/prov-overview>
- RDF: Resource Description Framework <https://www.w3.org/RDF>
- SKOS: Simple Knowledge Organization System <https://www.w3.org/2004/02/skos>
- SKOSXL: Simple Knowledge Organization System – eXtension for Labels <https://www.w3.org/TR/skos-reference/skos-xl.html>
- TBX: Term Base Exchange <http://www.ttt.org/tbx/>
- TEI: Text Encoding Initiative <http://www.tei-c.org/index.xml>
- URI: Uniform Resource Identifier <https://www.w3.org/Addressing/URL/uri-spec.html>
- URL: Uniform Resource Locator <https://url.spec.whatwg.org>
- VIAF: Virtual International Authority File <https://viaf.org>
- WGS: World Geodetic System, schema to specify geographic and cartographic information
- XSD: Schema Definition Language <https://www.w3.org/TR/xmlschema11-1>

## References

- [Alexiev 2015] Vladimir Alexiev. [GLAMs Working with Wikidata](#). In Europeana Food and Drink content provider workshop, Athens, Greece, May 2015.
- [Alexiev 2015b] Vladimir Alexiev. [Name Data Sources for Semantic Enrichment](#). Part of Europeana Creative Deliverable D2.4, Europeana Creative project, February 2015.
- [Alexiev 2015c] Vladimir Alexiev, [Thoughts on EHRI1 Linked Open Data](#) (Internal document re T11.3), July 2015
- [Gertner et al 2015] Thesaurus translated in other languages. EHRI deliverable 18.3, January 2015
- [Manske 2014] Magnus Manske, Wikidata [Mix-n-Match](#). 2014-2016
- [Ontotext 2015] [WikiProject Authority control](#), 2015
- [Orlandi et al 2014] Silvia Orlandi, Raffaella Santucci, Vittore Casarosa, Pietro Maria Liuzzo (editors). Information Technologies for Epigraphy and Cultural Heritage. Proceedings of the First EAGLE International Conference, 2014
- [PACKED 2015a] [Flemish art collections, Wikidata and Linked Open Data](#). PACKED vzw, Flemish Centre of Expertise in Digital Heritage, 2015
- [PACKED 2015b] [Flemish art collections, Wikidata and LOD: Whitepaper](#). PACKED vzw, Flemish Centre of Expertise in Digital Heritage, 2015
- [Wikidata 2015a] [WikiProject sum of all paintings](#), 2015
- [Wikidata 2015b] [Sum of all paintings: Benefits for museums](#), 2015
- [Mazzini, Ricci 2011] Silvia Mazzini, Francesca Ricci, [EAC-CPF Ontology and Linked Archival Data](#). Proceeding of the 1st International Workshop on Semantic Digital Archives, 2011
- [Eito-Brun, 2014] Ricardo Eito-Brun, [Remote access to EAC-CPF context and authority records for metadata indexing: a solution based on open information retrieval standards](#). Archival Science, December 2014
- [Lopez and Romary, 2010] [GRISP: A Massive Multilingual Terminological Database for Scientific and Technical Domains](#). In Proceedings LREC 2010. Malta.