Implementing an archival, multi-lingual and Semantic Web-compliant taxonomy by means of SKOS (Simple Knowledge Or-ganization System)

Francesco Gelati
Institut für Zeitgeschichte München — Berlin
Leibniz Institute for Contemporary History
Munich, Germany
gelati@ifz-muenchen.de

Abstract

This paper shows how a multilingual hierarchical thesaurus, or taxonomy, can be created and implemented in compliance with Semantic Web requirements by means of the data model SKOS (Simple Knowledge Organization System). It takes the EHRI (European Holocaust Research Infrastructure) portal as an example, and shows how open-source software like *SKOS Play!* can facilitate the task.

1 Introduction

Research projects, cultural heritage institutions, online repositories and catalogues often develop their own controlled vocabulary, which are primarily utilised as keywords in order to provide a thematic access to their entries. A catalogue entry that for instance displays "Refugee organisations" and "Relief and welfare organisations" in the metadata field "subjects" will be directly findable by users interested in these topics. This will be however possible so long as the users can browse the list of possible keywords, and perform keyword-based queries. In this paper I shall focus on a specific type of controlled vocabulary: taxonomies. Even though the term "taxonomy" is rarely used in human and social sciences, it is indeed the best option in order to describe hierarchical-structured controlled vocabulary, in the cultural heritage sector too. Cultural institutions are progressively sharing their catalogue entries, may they be archival descriptions, bibliographical records, museum data or digital objects, according to the FAIR principles. The same cannot unfortunately be said about their underlying taxonomies, which are simply not made exportable and reusable. Taking the EHRI (European Holocaust Research Infrastructure) Portal[1] as an example, I shall show in this paper how a taxonomy can be enriched with multilingual values and made interoperable by means of the semantic web data model SKOS[2] (Simple Knowledge Organization System) recommended by the W3C (World Wide Web Consortium), based on the RDF (Resource Description Framework) and compatible with the international standard ISO 25964-1 — Thesauri for Information Retrieval.

Some previous works: (Gelati, 2019), (Smith, 2018), (Vanden Daelen et al., 2015).

2 From Keywords to Taxonomies

The EHRI (European Holocaust Research Infrastructure) portal aims to aggregate digitally available archival descriptions concerning the Holocaust. This portal is actually a meta-catalogue, or an information aggregator, which imports datasets from a variety of data providers. Imported archival descriptions very often include the field "subjects" which bears keywords from the data provider's controlled vocabulary. Both archival descriptions and their keywords are written in many languages. In order to make keywords written in different languages equally findable, cross-lingual reconciliation is necessary. This concretely means that the English keyword "Refugee organisations" needs to be associated with its equivalent terms in all other supported languages (e.g. "organizacje uchodźców" in Polish). This is way EHRI developed a multilingual Holocaust and antisemitism-related taxonomy starting from previous hierarchies already used by partner institutions. The taxonomy was then made SKOS-compliant. Let us take a closer look to the SKOS specifications.

2.1 Concepts, Labels and Other Properties

"Concepts" are SKOS's main feature. Concepts

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

[2] https://www.w3.org/2004/02/skos/

Proceedings of the Workshop on Language Technology for Digital Historical Archives in conjuction with RANLP-2019, pages 24–27,
Varna, Bulgaria, Sep 5, 2019.
http://doi.org/10.26615/978-954-452-059-5_005

> "are identified with URIs, labeled with strings in one or more natural languages, documented with various types of note, semantically related to each other in [...] hierarchies and association networks, and aggregated into concept schemes".[3]

In our case, the EHRI taxonomy itself, called "EHRI terms"[4] is the concept scheme that incorporates all the concepts. Each term of the taxonomy (e.g. "Refugee organisations") is a concept, which is indeed provided with a URI, e.g.

```
https://portal.ehri-project.eu/
        keywords/ehri_terms-1199
```

and which can be expressed in all the natural languages[5] we wish by means of labels. Three types of label can be used: "preferred label", "alternative label" and "hidden label". In order to have a brief overview[6] of the rules, please note that, in order to avoid clashes, each concept may have no more than one preferred label for each language. The same value may not be used twice as preferred and as alternative value (nor twice as preferred and hidden, nor twice as alternative and hidden). Each concept may have as many preferred, alternative and hidden labels as wished. None of the three types of label is obligatory: a concept may have no labels at all, or may have for instance alternative label(s) only. Some of the above-mentioned concept "Refugee organisa-tions" preferred labels result as:

```
skos:prefLabel "Refugee
 or-ganisations"@en,
 "Flüchtlingsor-ganisation"@de,
 "organizacje uchodźców"@pl .
```

The fields "scope note", "definition" and "notation" may provide additional information or explanation on the concept.

```
skos:scopeNote "Refers both
 to refugees and to
 asylum-seekers
 organisations."@en .[7]
```

The properties "narrower" and "broader" shape the hierarchical tree of the taxonomy. In our case, the concept "Refugee organisations" has two broader concepts, "refugees" and "organisations", whose URIs are expressed below.

```
skos:broader
   <https://portal.ehri-project.eu/
   keywords/ehri_terms-1196> ,
  <https://portal.ehri-project.eu/
   keywords/ehri_terms-304> .
```

The possibility to create associative (i.e., non hierarchical) links between two or more concepts is also provided by the property "related". Some more properties, the most important being the "class" option, are equally available.

2.2 Multilingualism

Multilingualism is a strong feature of the EHRI taxonomy, for the following languages are implemented: Czech, Dutch, English, French, German, Hebrew, Hungarian, Italian, Polish, Russian, Serbo-Croatian and Ukrainian. They encompass three scripts, i.e. Latin, Hebrew and Cyrillic. Hebrew is displayed in Hebrew characters only, Serbo-Croatian in Latin only, whereas Russian and Ukrainian are parallelly displayed both in Latin and in Cyrillic:

```
skos:altLabel "Організації
   допомоги біженцям"@uk ,
   "Organìzacìï dopomogi
   bìžencâm"@uk-Latn ;
skos:prefLabel "organizacii
   bežencev"@ru-Latn ,
   "организации беженцев"@ru .
```

The taxonomy can updated by uploading to the online catalogue a new version of the taxonomy as a SKOS-compliant turtle file (.ttl). It means that new or amended concepts and their labels can be introduced at any time. So can always new languages be implemented. A user-friendly and code-free option for managing an existing taxonomy, or creating a new one, would be the web-based open-source tool "SKOS Play!"[8].

2.3 Creating a New Taxonomy

SKOS Play! provides you with a sample Excel spreadsheet, where each column relates to a given

[3] https://www.w3.org/TR/skos-primer/

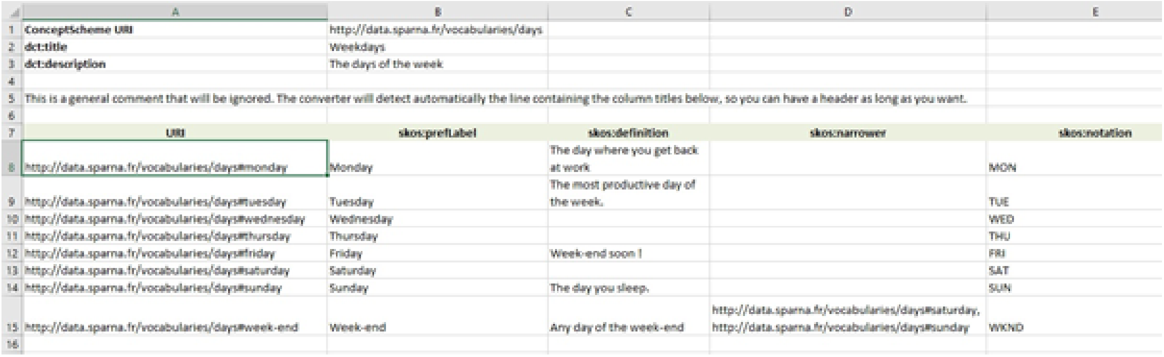
[4] https://portal.ehri-project.eu/vocabularies/ehri_terms/

[5] Hereafter will "language" or "natural language" always refer to the combination of a natural language and a script. It means that in this paper, simply for conciseness rather than for scientific purposes, Ukrainian written in Latin characters and Ukrainian in the Cyrillic script are considered two different languages.

[6] Please refer to https://www.w3.org/TR/2009/REC-skos-reference-20090818/ which is however at the moment of writing (2019-06-27) still a draft.

[7] Sample invented by the author.

[8] See: http://labs.sparna.fr/skos-play/. SKOS Play! is an open-source application developed by Thomas Francart for Sparna and released at the moment of writing under the licence CC-BY-SA 3.0.



	A	B	C	D	E
1	**ConceptScheme URI**	http://data.sparna.fr/vocabularies/days			
2	**dct:title**	Weekdays			
3	**dct:description**	The days of the week			
4					
5	This is a general comment that will be ignored. The converter will detect automatically the line containing the column titles below, so you can have a header as long as you want.				
6					
7	**URI**	**skos:prefLabel**	**skos:definition**	**skos:narrower**	**skos:notation**
8	http://data.sparna.fr/vocabularies/days#monday	Monday	The day where you get back at work		MON
9	http://data.sparna.fr/vocabularies/days#tuesday	Tuesday	The most productive day of the week.		TUE
10	http://data.sparna.fr/vocabularies/days#wednesday	Wednesday			WED
11	http://data.sparna.fr/vocabularies/days#thursday	Thursday			THU
12	http://data.sparna.fr/vocabularies/days#friday	Friday	Week-end soon !		FRI
13	http://data.sparna.fr/vocabularies/days#saturday	Saturday			SAT
14	http://data.sparna.fr/vocabularies/days#sunday	Sunday	The day you sleep.		SUN
15	http://data.sparna.fr/vocabularies/days#week-end	Week-end	Any day of the week-end	http://data.sparna.fr/vocabularies/days#saturday, http://data.sparna.fr/vocabularies/days#sunday	WKND
16					

Figure 1: The *Skos Play!* Excel file sample.

property (e.g. preferred label), and each row to one single item (e.g. a concept).

You can download the spreadsheet and enter there your own values. Then you can upload it back to the tool and convert it from Excel to a Semantic-Web and SKOS-compliant turtle file (.ttl), which will look like:

```
@prefix skos:  <http://www.w3.org/
     2004/02/skos/core#> .
<https://portal.ehri-project.eu/
     keywords/ehri_terms-989/>
     a skos:Concept ;
skos:prefLabel “Fascist
       propaganda”@en.
<https://portal.ehri-project.eu/
     keywords/ehri_terms-342/>
     a skos:Concept ;
skos:prefLabel “Antisemitic
       propaganda”@en.
<https://portal.ehri-project.eu/
     keywords/ehri_terms-986/>
     a skos:Concept ;
skos:prefLabel “Propaganda”@en ;
skos:narrower
     <https://portal.ehri-project.eu/
        keywords/ehri_terms-342/> ;
skos:narrower
     <https://portal.ehri-project.eu/
       keywords/ehri_terms-989/>.
```

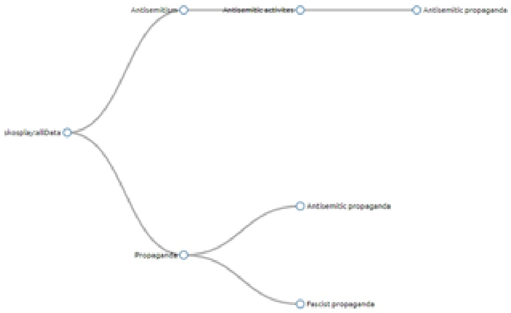


Figure 2: A *Skos Play!* visualisation option .

You may also visualise the data in a variety of options, amongst others as a tree.

2.4 Data Enrichment

Assigning URIs to all the entries of a digitally-shared taxonomy has many benefits. It permits first of all data enrichment from Linked Open Data multilingual databases like Wikidata.

One can automatically reconciliate identical entities, e.g. by means of the open-source programme Open-Refine.[9] One will then be able to associate the EHRI taxonomy entry “Propaganda” with its similar entry in Wikidata:

```
https://portal.ehri-project.eu/
     keywords/ehri_terms-986
=

https://www.wikidata.org/wiki/Q7281 .
```

It is also possible to manually create our own RDF triples, the standard way to make machine-readable affirmations. “Antisemitic propaganda is a category of Propaganda” may be expressed by means of the Wikidata property “subclass of”[10]: the former is a subclass of the latter will give

```
<https://portal.ehri-project.eu/
     keywords/ehri_terms-342/>
<https://www.wikidata.org/wiki/
     Property:P279>
<https://portal.ehri-project.eu/

     keywords/ehri_terms-986/> .
```

3 Conclusion

By means of the few steps described above, an online archival (meta)catalogue can make its multilingual taxonomy digitally available and machine-readable. The possibility to manage a SKOS taxonomy in a variety of formats (including TTL, RDF/XML and JSON), attribution of URIs (which the research body simply has to activate), linkage of information with leading open-source databases, compliancy with Semantic-Web

[9] http://openrefine.org/

[10] Whose URI is: https://www.wikidata.org/wiki/Property:P279

requirements… Everything makes the data FAIR: findable, accessible, interoperable and reusable.

Acknowledgments

The author, who was introduced by Ksenia Zaytseva and the Austrian Centre for Digital Humanities to SKOS, wishes to thank both of them, as well as the EHRI staff.

References

Francesco Gelati. 2019. Archival Metadata Import Strategies in EHRI. *ABB: Archives et Bibliothèques de Belgique - Archief- en Bibliotheekwezen in België*, Trust and Understanding: the value of metadata in a digitally joined-up world, ed. by R. Depoortere, T. Gheldof, D. Styven and J. Van Der Eycken(106):15–22.

Jeffrey Smith. 2018. Toward “Big Data” in Museum Provenance. In Giovanni Schiuma and Daniela Carlucci, editors, *Big Data in the Arts and Humanities. Theory and Practice*, pages 41–50. Taylor & Francis, Boca Raton, FL.

Veerle Vanden Daelen, Jennifer Edmond, Petra Links, Mike Priddy, Linda Reijnhoudt, Václav Tollar, and Annelies van Nispen. 2015. Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives. In *“Open History: Sustainable digital publishing of archival catalogues of twentieth-century history archives”*, Brussels, Belgium.