A COMMON SSH VOCABULARY INFRASTRUCTURE?

Daan Broeder, CLARIN ERIC

2022 CLARIN centre days
Vocabularies in the SSHOC project

• Coordination wrt. vocabularies: originally a limited effort
  • Investigation of a common recommended platform for publishing and sharing vocabularies
  • Testing machine translation for vocabularies
  • Flexible integration of vocabularies in tools: e.g. SSHOC Dataverse

• Identified more opportunities during the project
  • Training & discussion: CLARIN vocabulary initiative 2020
    https://www.clarin.eu/blog/clarin-sshoc-vocabulary-initiative
  • Recommendations for further common approaches e.g. Vocabulary versioning, investigating CV authoring tools
  • Vocabulary recommendations for researchers
  • Opportunity & need to represent SSH interests with other stakeholders e.g. software & service providers
SKOSMOS

• SKOSMOS was selected as the recommended software for vocabulary publication

• Already used by some DARIAH, CESSDA centers and beyond the SSH, good relations with solid development team based at NFL, fulfills most technical requirements: API, supported formats, …
  • But note SKOSMOS is not an authoring tool,
  • its targeted audience and performance do not cover all our use cases

See:
https://sshopencloud.eu/ms8-choice-vocabulary-publication-platform-sshoc
SKOSMOS

• Out of the box SKOSMOS is
  • Jena/fuseki triple database
  • PHP / JS based frontend
  • Supports different concept schemes in one vocabulary definition file and skos collections/groups
  • Frontend can be customized

• Can be used for large vocabularies:
  Agrovoc: 40k concepts, 750k triples (https://agrovoc.fao.org/browse/agrovoc/en/), 17m load M1max cpu, but need optimized triple store for responsive UI and querying

Examples of current users

• Finto
• BARTOC vocabularies
• CESSDA / ELLST thesaurus
• CNRS / Loterre, a multidisciplinary terminology platform
• FAO / AGROVOC
• Luxembourg Service central de législation Controlled vocabularies
• Rhineland-Palatinate spatial data initiative classifications
• UK Data Service / HASSET - The Humanities and Social Science Electronic Thesaurus
• UNESCO / UNESCO Thesaurus
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Vocabulary information

TITLE
CCR - CLARIN Concept Registry (v2016)

TYPE
http://www.w3.org/2004/02/skos/core#ConceptScheme

URI
http://hdl.handle.net/11459/CCR_P-DialogueActs_955814a6-6c07-c143-94eb-b2551c2d51cb

Resource counts by type

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept</td>
<td>3079</td>
</tr>
<tr>
<td>Collection</td>
<td>102</td>
</tr>
</tbody>
</table>

Term counts by language

<table>
<thead>
<tr>
<th>Language</th>
<th>Preferred terms</th>
<th>Alternate terms</th>
<th>Hidden terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>3079</td>
<td>176</td>
<td>11</td>
</tr>
</tbody>
</table>
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782
<table>
<thead>
<tr>
<th>Vocabulary information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TITLE</strong></td>
</tr>
<tr>
<td><strong>LAST MODIFIED</strong></td>
</tr>
<tr>
<td><strong>TYPE</strong></td>
</tr>
<tr>
<td><strong>VOID:INDATASET</strong></td>
</tr>
<tr>
<td><strong>URI</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resource counts by type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
</tr>
<tr>
<td>Concept</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term counts by language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
</tr>
<tr>
<td>Arabic</td>
</tr>
<tr>
<td>Catalan</td>
</tr>
<tr>
<td>Czech</td>
</tr>
<tr>
<td>Danish</td>
</tr>
<tr>
<td>German</td>
</tr>
<tr>
<td>Greek</td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td>Spanish</td>
</tr>
<tr>
<td>Estonian</td>
</tr>
<tr>
<td>Persian</td>
</tr>
<tr>
<td>Finnish</td>
</tr>
<tr>
<td>French</td>
</tr>
<tr>
<td>Hindi</td>
</tr>
</tbody>
</table>
But wrt. vocabulary discovery & researcher recommendations

- SKOSMOS does not provide useful registry functionality
- Only a scrollable list of vocabularies displaying vocabulary metadata
- The UI offers no metadata search, no facets, no keyword search, although for the content (terms) extensive search & filter options are offered
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782.

Each concept in YSO is issued to at least one thematic group and some concepts have also been grouped with a non-hierarchical collection label (such as “clothes by materials”).

Following the international standards for thesauri, the terms for concepts are usually plural nouns. Terms in singular are usually mass nouns or terms referring to actions or abstract concepts. Some terms carry a different meaning when used in plural and in singular. For example, ballet refers to an art form and ballerinas to individual works of art.

Place names are contained in a separate ontology, YSO places.

YSO is based on General Finnish Thesaurus (YSA) and General Finnish Thesaurus in Swedish (Allås). Concepts in YSO and concepts in YSA and Allås have been linked to each other with equivalence relationships. YSO has also been linked to Library of Congress Subject Headings (LCSH).

<table>
<thead>
<tr>
<th>PUBLISHER</th>
<th>National Library of Finland</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREATOR</td>
<td>National Library of Finland Semantic Computing Research Group (SeCo) The Finnish Terminology Centre TSK</td>
</tr>
<tr>
<td>SOURCE</td>
<td>YSO pohjautuu yleiseen suomalaiseen asiakasannostoon (YSA) sekä yleiseen ruotsinkieliseen tesauroikeeseen (Allås).</td>
</tr>
<tr>
<td>LICENSE</td>
<td><a href="http://creativecommons.org/licenses/by/4.0/">http://creativecommons.org/licenses/by/4.0/</a></td>
</tr>
<tr>
<td>LAST MODIFIED</td>
<td>Friday, December 18, 2020 04:43:58</td>
</tr>
<tr>
<td>TYPE</td>
<td><a href="http://www.w3.org/2004/02/skos/core#ConceptScheme">http://www.w3.org/2004/02/skos/core#ConceptScheme</a></td>
</tr>
<tr>
<td>URI</td>
<td><a href="http://www.yso.fi/onto/ysos/">http://www.yso.fi/onto/ysos/</a></td>
</tr>
</tbody>
</table>

Download this vocabulary: TURTLE
Vocabulary visibility and discovery

• Vocabularies not always FAIR yet; they should be properly registered and published, researchers & infrastructure providers should be able to find and reuse -> FAIR semantic artefacts

• SSH Vocabulary registry or a general one that supports sufficient discipline specificity e.g. Bartoc (3300 entries whereof 1200 SSH)

• Vocabulary search facility, that searches in vocabulary metadata but also the vocabulary terms themselves.

• Note that providing optimal recommendations for researchers can be complicated e.g. also aspects of context and user profile play a role

Bartoc registry: https://bartoc.org/vocabularies
FAIR Vocabularies - tool access to vocabularies via APIs

Two strategies

• Every tool knows multiple APIs
  • Scalable approach?
• Intermediate broker allows also caching vocabularies?
  • Added maintenance
  • Caching seems only useful in case of unreliable providers
Towards a SSH Vocabulary Commons

Common interest by RIs: CESSDA, CLARIN, DARIAH and E-RIHs

Collaborative use and management of vocabularies
vocabularies as first-class citizens / FAIR data objects in their own right.

Mission: vocabulary commons charter

Include other projects initiatives EU eg. TRIPLE and national
Priorities for the Vocabulary Commons

• Operating a Vocabulary repository
• SSH vocabulary overview
  • of all relevant SSH vocabularies
• Vocabulary federated content search
  • “Deep” search also in the terms (essential for recommendations)
• Harmonized versioning of vocabularies
  • yet no agreed way of how to manage update of vocabularies and individual vocabulary terms
• API standardisation:
  • to enable tools to interact smoothly with different vocabulary platforms
• Foster exchange between users and developers (eg SKOSMOS)
SKOSMOS instance to publish the SSHOC vocabulary results and other SSH vocabularies

https://vocabs.sshopencloud.eu/vocabularies/
Vocabulary visibility and discovery

- Vocabularies not always FAIR yet; they need proper registration and publication so researchers & infrastructure providers can discover and reuse -> see the FAIRsFAIR project report for FAIR semantic artefacts
- Need a SSH Vocabulary catalogue or a general one that supports sufficient discipline specificity e.g. Bartoc.org (3300 entries whereof 1200 SSH)
- Vocabulary search function, that searches in vocabulary metadata but also the vocabulary terms themselves.
- Note that providing optimal recommendations for researchers can be complicated e.g. also aspects of context and user profile play a role
Vocabulary Search

Effective vocabulary recommendations for researchers require:

- Browsing/searching via metadata AND
- Searching in the vocabulary terms

Considering two strategies:

- Central metadata and content harvesting and indexing (update/syncing problem)
- Federated search, possibly on the basis of SKOSMOS instances and SPARQL endpoints (performance problem)
Collaborative Vocabulary Management

Vocabulary management is “solved” and reuse by means of copy too

BUT:

Reuse and management of a vocabulary across multiple organisational scopes not

- Ownership in a distributed (loosely coupled) setup => authority
- Procedures for evolution and agreement
  (new concepts, semantic shift)
- Synchronisation with source/target applications
  Ideally, user wants to adjust vocabulary (add new concepts) at point of use
Towards a SSH Vocabulary Commons

Common interest by RIs: CESSDA, CLARIN, DARIAH and E-RIHs

Collaborative use and management of vocabularies as first-class citizens / FAIR data objects in their own right.

Mission: vocabulary commons charter

SSHOC activities wrt vocabularies:

- tasks 3.1 Multilingual terminologies, 3.5 Interoperability, WP7 SSH Open Marketplace usage of vocabularies etc.
- CLARIN Vocabulary Initiative inventorizing relevant SSH vocabularies and organizing info sessions on vocabulary management software and SSH requirements
- ICTeSSH workshop on vocabulary use in the SSH
- Wider scope for collaboration in the Humanities eg. TRIPLE project
Collaborative Vocabulary Management

Vocabulary management is “solved” and reuse by means of copy too

BUT NOT

Reuse and management of a vocabulary across multiple organisational scopes

• Ownership in a distributed (loosely coupled) setup => authority
• Procedures for evolution and agreement
  (new concepts, semantic shift)
• Synchronisation with source/target applications
  Ideally, user wants to adjust vocabulary (add new concepts) at point of use

Note also that changes in vocabularies should be made explicit and documented beyond versioning eg.
Note: In the CMDI TF meeting tomorrow Menzo will present what happens with regard to vocabularies in CLARIAH-NL

Thank you for your attention
Vocabularies & Interoperability

• **Technical / Format interoperability.** SKOS and OWL are broadly accepted
  • but many projects use spreadsheets and tables and are locked in silos using highly specific software to manage and use these
  • Specific recommendations for vocabulary versioning are needed

• **Semantic interoperability.** Coming from different traditions different organizations and projects have developed different vocabularies to describe similar data. Normalization or conversion needed; the vocabularies involved can be huge and expertise expensive, good tools exist (Ariadne VMT)

• **Organizational interop.** when sharing also responsibility for vocabulary maintenance, what model for (non-)agreement can we have

• **Cultural & Human interop** aspects. Multilingual vocabularies, localization aspects.
Vocabularies & Interoperability

For data reuse and data integration we have to look at interoperability of vocabularies

- **Technical / Format interoperability.** SKOS and OWL are broadly accepted
  - but many projects use spreadsheets and tables and are locked in silos using highly specific software to manage and use these
  - Specific recommendations for vocabulary versioning are needed

- **Semantic interoperability.** Coming from different traditions different organizations and projects have developed different vocabularies to describe similar data. Normalization or conversion needed; the vocabularies involved can be huge and expertise expensive (Ariadne Vocabulary Matching Tool).

- **Cultural & Human interoperability** aspects. Multilingual vocabularies, localization aspects. -> MT technology + network of human experts