Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782 Start Date of Project: 01/01/2019 Duration: 40 months

Report on Milestone **MS8.0**

Choice of Vocabulary Publication platform for SSHOC

Dissemination Level	PU
Due Date of Milestone	30/06/2019, M8
Actual Achievement Date	**30/06/2019**
Lead Beneficiary/LTP	16. CNR
Work Package	WP3 Lifting Technologies and Services into the SSH Cloud
Task	Task 3.1. Multilingual Terminology
Version	V1.2
Date	11/06/2021
Number of Pages	p. 1 – p. 36

Abstract: The goal of the activities in SSHOC Task 3.1. is to permit metadata-based discovery in different languages and one of the main tasks is to propose the SSHOC vocabulary publication platform. To achieve this goal, a survey and interviews were first conducted with users of vocabulary publication platforms to identify the core features needed for the editing, linking and publication of vocabularies. The results were then used to evaluate the existing vocabulary publication platforms (e.g. CESSDA, SKOSMOS etc.) that allow the discovery of the research data in SSH infrastructures and to provide recommendations for a suitable vocabulary publication platform.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability. This deliverable is licensed under a Creative Commons Attribution 4.0 International License.

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

Author List

Organisation	Name	Contact Information
CNR(ILC) and CLARIN-IT	Monica Monachini	monica.monachini@ilc.cnr.it
CESSDA/FSD	Taina Jääskeläinen	taina.jaaskelainen@tuni.fi
CLARIN	Dieter Van Uytvanck Iulianna Van der Lek Daan Broeder	dieter@clarin.eu i.vanderlek@uu.nl daan.broeder@di.huc.knaw.nl
DARIAH	Yoann Moranville	yoann.moranville@dariah.eu

Contents

1. Introduction

It is important to understand the general framework of use for a vocabulary publication platform for the SSHOC project. In SSHOC, it is crucial to support better discovery of SSH research data in order to ensure better access and reusability. This will be made possible through support for multilinguality. In infrastructures, metadata aggregation platforms are provided that map metadata to a shared common ontology usually in English. A vocabulary server and publication platform will be provided in SSHOC to maximize the accessibility and to improve discovery of content by non-native speakers, thus allowing multilinguality. To achieve this, this Milestone reports on:

(a) a survey of existing systems for managing (editing and publishing) and accessing (browsing) vocabularies used to describe and allow discovery of research data in SSH infrastructures; the systems are described based on relevant features:
   - https://docs.google.com/spreadsheets/d/1MV2g1PZMQ_Rx8m8tybthOY9eMuFNpdJxUCfvu-jxuLI/edit#gid=0

(b) a series of interviews with experts to identify the core features of vocabularies and vocabulary platforms and what was missing:
   - https://docs.google.com/spreadsheets/d/1FSDlEPBfYdRU6TQgPXoWofpibkAGGv3YzZrMy0FBsIo/edit#gid=0

The two documents[1] have been cross-compared to extract a list of relevant criteria and characteristics that the SSHOC publishing platform should have to support the editing, linking and publishing of the vocabularies. In addition, the following technical criteria have been identified and compared, such as the installation requirements, availability of an API and source code, implementation of Linked Data, and license type. The feature comparison table is available at:

- https://docs.google.com/spreadsheets/d/1s5_StggMB1AburKRG4U6kK6oj5Dsc4weaGtXw0w1YNA/edit#gid=0[2]

[1] Provided links to the documents might have access restrictions as they are intended for the SSHOC Consortium partners and are stored in the internal SSHOC document repository which is hosted by project coordinator's Google Shared Drive and will be archived after project completion. For this reason, images of the documents are provided in the Appendices to this report.

[2] Provided links to the document might have access restrictions as it is primarly intended for the SSHOC Consortium partners and stored in the internal SSHOC document repository which is hosted by project coordinator's Google

The support documents (1. survey of existing systems – 2. series of interviews with experts – 3. set of criteria for comparison) are also provided in the Appendices to this report.

Note that in version 1. 2 of this document, new information was added wrt. the current situation (Q1 2021) for the use of vocabulary platforms in the CLARIN infrastructure and the results of the CLARIN Vocabulary initiative actions, which were added as chapter 4 and appendices. New information in the existing chapters was added in Italics.

2. Description of the Milestone

a. Role of the Milestone

This section provides a review of the most used vocabulary management platforms to identify the core features that support the creation, maintenance and publication of vocabularies. The following systems have been briefly evaluated:

- The ACDH Vocabulary Service
- The Thesaurus Management System – THEMAS
- VocBench3
- The CESSDA Vocabulary Service
- OpenSkos
- iQvoc

Review of vocabulary management platforms

The **ACDH Vocabulary Service (Vocabs)** provides services and tools that allow for collaborative creation, maintenance and publication of vocabularies and taxonomies of any kind. Vocabs is a vocabulary repository service that enables browsing of vocabularies with structured concept displays and visualisation of concept hierarchies. Vocabularies can be searched with a search interface or by consulting an alphabetical or thematic index. Vocabularies can be accessed via a REST-API to allow for Linked Data.

The Vocabs editor follows the SKOS data model for the main elements of a vocabulary. The Dublin Core schema is used to capture the metadata (such as date created, date modified, creator, contributor,

Shared Drive and will be archived after project completion. For this reason, image of the document is provided in the Appendices to this report.

source and other) about each element. Each concept scheme (vocabulary/ontology/thesaurus) can be downloaded in RDF/XML and Turtle format as well as each individual concept. The user management system allows a user to share a concept scheme they created with other users (called 'curators') to create new concepts, edit and delete concepts and collections within this concept scheme. Each user can find a summary of their latest activity on the user's page.

The system is based on the open-source software Skosmos, which uses SKOS as the underlying data model. Skosmos enables users to browse the vocabularies with structured concept displays and visualisation of concept hierarchies. Vocabularies can be searched with a search interface or by consulting an alphabetical or thematic index. Vocabularies can be accessed via a REST-API, to allow for Linked Data.

The Vocabulary Service does not only serve controlled vocabularies for ACDH purposes but is also employed for some of the central services of the CLARIN infrastructure: the CLARIN Concept Registry and CLAVAS, a repository for specific metadata vocabularies. Additionally, ACDH aims to offer Vocabulary Service as a service for partners of the national CLARIAH-AT consortium, for the Thesaurus Maintenance working group in DARIAH-EU, and PARTHENOS.

The **Thesaurus Management System – THEMAS** is an open-source Web-based system for creating, managing and administering multi-faceted multilingual thesauri according to the principles of ISO 25964-1 and ISO 25964-2 standards.

The distinct features of the system include management and administration of semantic relationship types within thesaurus concepts, ease of navigation among interconnected terms, extensive search capabilities, multiple presentation displays of concepts and their context and more. THEMAS can be adjusted to fit the needs of any domain of research, through customizable system configuration options for the activation/deactivation of integrated consistency controls, the user interface language, the thesaurus dominant language, the set of translation languages, the customization of graphical representation and so on. The workflow control offered by the system harnesses well-defined user roles to facilitate the orderly evolution of the managed thesauri. It allows for the collection of knowledge from a large number of users, while smaller groups of users with domain expertise in the target thesaurus can be assigned responsibility for the approval, correction or rejection of any contribution, thus ensuring data correctness and thesauri consistency. Thesauri data – either as a whole thesaurus or as a result of an arbitrary search – are made available to users online and all thesauri data are exportable in machine-readable form to enable their easy integration into diverse information systems deploying thesauri. The system provides XML import, export, and SKOS export, and can also provide on-demand SKOS import. The internal schema used is generally more detailed and flexible than SKOS. The software itself is offered as an Open Source solution. The THEMAS user interface is implemented as a web application using Java Servlets technology. Data storage and maintenance is performed using the open-source graph database Neo4j community edition, while the thesauri structures follow the principles of TELOS representation language. The web application communicates with the data storage via Neo4j-sisapi. All related

technologies involved are Java-based, thus making the system installation compatible with any platforms[5].

The **VocBench3** is an open-source web application for editing thesauri complying with the SKOS and SKOS-XL standards. VocBench has a strong focus on collaboration, supported by workflow management for content validation and publication. VocBench adopts a role-based access control mechanism, checking user privileges for requested functionalities through the role they assume. The presentation layer is implemented as a Web application, powered by GWT, and by a series of lightweight Web services. The business and the data layers are both implemented using the Semantic Turkey RDF management platform [4] extended with ad-hoc software programs [3]. The current release, VB3, offers a flexible environment, fine-grained RDF editing, support for several core modelling vocabularies. It also offers a powerful editing environment, with facilities for the management of OWL ontologies, SKOS/SKOS-XL thesauri, OntoLex lexicons and any sort of RDF dataset. The development of VocBench is managed by the Publications Office of the European Union, with the aim to help public administrations to maintain and publish their controlled vocabularies in an open and interoperable way.

The Editor is integrated into the browser, not very nice looking or user-friendly but seemed to be functional. No-code custom forms can be created in v5. Stylesheet and it is easy to alter. It supports both project-based, as well as customised roles. Import and export of vocabularies are possible via GUI in multiple serialisations. Native SKOS class, and SKOS class attributes and note types. The platform supports around two hundred languages on all properties. *For CESSDA, VocBench is now the choice for ELSST thesaurus editor and SKOSMOS as publication platform.*

The **CESSDA Vocabulary Service** is a web application that enables users to discover, browse, and download controlled vocabularies in a variety of languages. The service has a web Editor as well as a browsing and download interface. The tool is provided by the Consortium of European Social Science Data Archives (CESSDA). It has an API and it is open-source with a Bitbucket repository. The majority of the source (English) vocabularies included in the service have been created by the DDI Alliance. The Data Documentation Initiative (DDI) is an international standard for describing data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences. The language versions of the DDI vocabularies have been provided by CESSDA member organisations or other organisations, as the vocabularies are used in metadata to describe research data. The Editor provides user management. Only authorised users can create, manage and translate vocabularies. Roles are language- and agency-specific. One user can have more than one role. Versioning is at the vocabulary level, each language version is versioned independently, and versioning only applies to published versions. Translated versions, however, are connected to a particular source version, and can only begin after the source version has been published. Translators need to translate the whole vocabulary before publishing it. The system enforces the translation of all descriptive terms of concepts. Definition translation is not mandatory. Once a new version of the source is published, all language versions of the previous versions are frozen and cannot be edited anymore. The systems clone the previous versions as

a starting point for the new version creation. At present, the system allows definitions and hierarchical relationships, but no synonyms or associative relationships.

OpenSKOS is a web platform providing a web service-based approach to publication, management and use of vocabulary data that can be mapped to SKOS[1]. The goal is to enable vocabulary producers to create, manage and share vocabularies. The OpenSKOS platform implements a distributed architecture based on the peer-to-peer paradigm, every peer (AKA OpenSKOS site) runs an instance of the OpenSKOS repository. Interactions between OpenSKOS peers and with other SKOS systems may occur at two levels: *Data level*, exchanging copies of vocabularies using the OAI-PMH protocol; *API level*, using a RESTful API embedded in every OpenSKOS site. The API also implements Linked Data functionalities. *In Q1 2021 it became apparent that the OpenSKOS platform will not be further developed, something that had consequences for the CLARIN infrastructure using OpenSKOS as its standard vocabulary platform*.

The **Skosmos vocabulary browser** is a controlled vocabulary-publishing tool that can manage multiple vocabularies and it is designed to be both user and developer-friendly. Using a three-layer architecture pattern, Skosmos can be described as follows: the presentation layer is composed of a Web GUI for users and a REST API for external agents, the business logic is composed of a set of tools used to create and publish vocabularies and the data layer is managed by a SPARQL endpoint[2]. According to the description in [2] the Web GUI enables users to browse and search the data and to visualize concept hierarchies, however, it does not provide management facilities that should be performed using SPARQL MS or other tools. Easy to install, steady development team. The main features are: search and browse vocabularies, alphabetical and thematic index, structured concept display, visualized concept hierarchy, multilingual user interface. Requires vocabularies to be SKOS. Skosmos provides a REST-style API and Linked Data access to the underlying vocabulary data. Current users include the Unesco thesaurus, Agrovoc thesaurus of FAO, Finto, and ACDH Vocabulary Service. Skosmos offers out-of-box features like alphabetical/hierarchical browsing, autocomplete search, URI-based content negotiation, and a feedback form. Important aspects for UNESCO were the ability to have a multilingual interface (English, French, Spanish, Russian), the possibility to customize the stylesheets/logo/help page, or the order of the fields in a concept display page[3]. The OpenSKOS software provides a Dashboard GUI that implements the management functionalities.

The **iQvoc** is a vocabulary management tool that supports the editing and publishing of SKOS datasets. It supports RBAC and for unauthenticated users, it provides search and browse access to concepts in the vocabulary. The platform is implemented using Ruby and Javascript libraries for the GUI and the business

[3] Example of SPARNA: http://blog.sparna.fr/2017/02/06/unesco-thesaurus-published-with-semantic-web-standards-and-open-source-software/ [accessed 30 Mar 2021]

logic, iQvoc uses a relational representation of the SKOS-XL model with an external RDF representation and SPARQL endpoint which are provided by Triple Store integration[6]. The system provides the following roles for registered participants of the editorial team: 1) 'editor' role for creating and modifying concepts but not for publishing them, 2) 'publisher' role for publishing and 3) admin for unlocking concepts, user management, export and import, set configuration. No separate translator or language-specific roles, so editors and publishers can amend all languages. Within a concept scheme/vocabulary, allows broader, narrower and top terms, related terms, translation of concepts and their definitions, and mapping to different concept schemes. Handling multilingual vocabularies seems a bit cumbersome. Each vocabulary/concept scheme needs to have its instance, with the specified editorial team and language choices.

General remarks about the features of a vocabulary platform in SSHOC

It is important to understand the general framework of use for the proposed vocabulary publication platform for the SSHOC project. In SSHOC it is crucial to support better discovery of SSH research data to ensure better access and reusability. This will be made possible through support for multilinguality. In infrastructures, metadata aggregation platforms are provided that map metadata to a shared common ontology usually in English. A vocabulary server and publication platform will be provided in SSHOC to maximize the accessibility and to improve discovery by non-native speakers, thus allowing multilinguality. The vocabularies will be available for download and API service.

Controlled vocabularies such as thesauri and classifications are published on the web for searching and browsing. Such vocabularies are published as RDF data using the Simple Knowledge Organization System (SKOS) model to represent concepts and their labels.

Publishing tools can be used to expose such vocabulary data on the web, but this is not enough for SSHOC users. Publishing tools specialized for SKOS vocabularies can provide search, browsing and other features specific to vocabularies, in addition to basic Linked Data publishing.

Need for editing functionalities

One of the major questions concerns the editing service. Is the common platform only going to be a point of access or does it need to offer editing/mapping services? Reviewed existing vocabularies already have their own editing facilities. Such services will continue to be used for different projects. They may also be a reference to manage other vocabularies (from other projects), but a common editing platform is one of the first requisites to consider since it allows for collaborative work on controlled vocabulary development that is important in SSHOC.

Need for linking/mapping/alignment functionalities

At the same time, in the context of research infrastructures, it is common to create metadata aggregation platforms. An alignment interface will be needed for reconciling, linking, mapping. The common editing

platform calls for functionalities to manage the process of vocabulary reconciliation. This platform hence seems to have strong requirements in terms of:

- Semi-automated alignment (suggestions, etc...)
- Synchronization
- Import/export
- Role management

We should consider whether systems carried out in previous mapping initiatives (PARTHENOS entities) can be used or integrated.

Need for a browsing interface

Reviewed vocabularies have their own publishing interfaces; a common user-friendly browsing interface will be needed. The common browsing interface will have to be publicly accessible for view only, and also offer multilingual (translated) vocabularies. In particular, it should be possible for everyone to retrieve the URIs of each concept and inspect various serialisations.

Size of the vocabularies

The surveyed vocabularies considerably vary in size:

- CCR 1520
- Tadirah & CESSDA (hundreds)

These numbers should be manageable for any existing system.

Requirements for a vocabulary platform in SSHOC

Based on the survey and interviews, we identify the following list of requirements for the vocabulary management platform to be used in SSHOC:

Data model:

- Skos seems to be the most used data model in all vocabularies
- The chosen platform needs to have skos I/O

Structure:

- All vocabularies are based on hierarchies of concepts
- Most surveyed vocabularies do not seem to implement related concepts or synonyms,
- The support of hierarchies of concepts and for such features may be an added value

Multilinguality**:**

- DARIAH vocabularies (TaDiRAH and Backbone) and CESSDA/DDI are multilingual (7 languages and increasing);
- Multilingual support is required to improve discovery by non-native speakers and maximize the accessibility

Concept schemes support:

- Some vocabularies such as CLARIN CR have concepts schemes.
- CS support is important in view of facilitating also faceted search

APIs support:

- API service is generally regarded to be an important feature for vocabulary access

Browsing interface:

- In some of the surveyed platforms, the browsing interface is separate from the editing interface.
- Based on the existing browsing interfaces, the following features have been identified:
    - Facet browsing
    - In particular browsing by source (agency/project/collection) is a useful feature in this context, as it might allow filtering concepts by provenance
    - A search interface with the possibility to organise the search results, e.g. alphabetically.
    - Basic tree view
    - Multilingual view and search
    - Autocomplete search, indexes preferred labels, alt labels and hidden labels

Web-based editing / advanced user management (roles):

- Based on the interviews, all projects are based on a system of roles for the maintenance and the incrementation of their vocabularies
- Mandatory feature for the chosen editor, given the necessity for various partners to collaborate in different capacities
- In particular, the possibility of managing institutions is crucial (Tenants/Agencies)
- One person should be able to have many roles; contributors should be differentiated from administrators and different languages should have different administrators.
- Producing translations would optimally be user-friendly: the system keeps the language choice without translators having to choose the language for every concept or concept element translated. Translators should see the original text to be translated all the time they are translating. They should be able to see other language versions easily.

Versioning and status:

- Strict versioning is a feature of the CESSDA Vocabulary Service where versioning takes place between published vocabularies. Version changes can be documented (though this is voluntary); some of the documentation is machine-generated. See the 'Versions' tab in this example[4]. There is also a comparison table. Strict versioning is used in CESSDA because organisations using the vocabularies in metadata may update their legacy metadata after changes in vocabularies, to keep filters based on them functional.
- It should be possible to go back to previous states of the vocabulary and mappings and reconstruct the history of modifications
- In addition, status and contributions should be monitored; in CLARIN CR, concepts have states. e.g., *candidate, approved* (wrt. the role of the contributors); in CESSDA Vocabulary Service, vocabularies themselves have status (draft, initial review, final review, published) and non-registered users can only see and browse the published vocabularies.

The requirements of the Vocabulary Management Platform can be summarized as follows:

- provide users with unified access to all vocabularies
- graphical interface for the management of multiple lists of vocabularies, taxonomies, thesauri;
- Management of terms should provide information about status, language, unique URI type identifier, notes of different types associated with terms, equivalence relations between concepts, hierarchical and associative relations between concepts, e.g. hierarchical relationships (narrower, broader), association relationships (related terms), semantic relationships (synonyms), authorised semantic relationships (one of the synonyms is identified as Preferred Term, PT), translations (and what elements are translated)
- management of facets;
- management of a set of concepts (micro-thesauri);
- import and export of thesauri in SKOS / RDF format;
- consultation through web services
- a web service-based editor for collaborative networking with collaborative functionalities
- friendly interface, suitable for use by non-expert users, coming from different communities
- logically structured and intuitive workflow

[4]Example of CESSDA Vocabulary Service, https://vocabularies.cessda.eu/#!detail/TopicClassification?url=https%3A%2F%2Fvocabularies.cessda.eu%2FTopicClassification%2Fen%2F3.0&lang=en, accessed 30 Mar 2021

- flexible enough to be adapted to new needs and standards
- alignment between vocabularies and external resources

b. Means of verification

According to the GA Table 1.3.4 WT4 List of milestones, MS8 is expected to provide the choice of the vocabulary publication platform for SSHOC. In the Report, we provide a survey of some of existing platforms and define a list of criteria and features against which the different existing platforms for the discovery of research data in SSH infrastructures are evaluated. This allows us to formulate motivated recommendations for SSHOC.

3. Preliminary Conclusions

From the analysis of surveyed editing platforms, it emerges that some of them already offer a wide range of solutions to meet the requirements of the SSHOC infrastructure.

CLARIN strongly relies on SKOS solutions. OpenSkos2 seems to address most of the requirements: Skos, Concept Schemes, Roles, versioning, API. The CLARIN Centre Committee has started in 2015 managing vocabularies in CLAVAS, the CLARIN Vocabulary Service. CLAVAS is a platform for SKOS vocabularies and has been established to provide centrally maintained (i.e. not open for editing) controlled vocabularies. In the future, the CLARIN Centre Committee would like to extend CLAVAS and allow it to be used as open vocabularies. Experiments with new ways to maintain and managing vocabularies in CLAVAS could be done in the context of SSHOC results. *Although OpenSKOS is still functional, in Q1 2021 CLARIN reevaluated the OpenSKOS technology support and decided to move to Skosmos as a vocabulary publication platform. Moving the current CLARIN supported vocabularies to the Skosmos platform and its integration with other CLARIN services is planned for 2021.*

The **CESSDA** Vocabulary Service offers an interesting solution including the browsing interface. The service doesn't seem best suited for large concept schemes with a great number of terms as the focus has been on vocabularies where potential users need to view the whole list.

The **ACDH** Vocabulary Service (used in DARIAH), has an Editor and uses Skosmos as the publication platform. Together, a suite of services that allow for collaborative creation, maintenance and publication of vocabularies and taxonomies of any kind.

After listing the requirements, we have also inspected the existing browsing interfaces for the various vocabularies:

The **CLARIN** interface provides faceted browsing, filter by project and concept scheme. However, it does not allow to inspect the source (directly or by means of appropriate serialization)

The **CESSDA** Vocabulary Service interface is similar. Filtering is possible by language and agency, and in the Editor, also by status. It has nice download features for export and multilingual view; it allows multilingual search, ordering the result list by relevance or alphabetically by vocabulary title.

The **ACDH** Vocabulary Service browser, based on SKOSMOS, also looks good, with a nice hierarchical tree view, alphabetical view and export in various serialisations (RDF/XML TURTLE JSON-LD); it also offers good support for multilingual view.

The result of the interviews is that not all the requirements of a full-fledged vocabulary publication platforms are covered in the solutions adopted by either CLARIN or CESSDA. Both CLARIN and CESSDA experts seem to be open to alternatives.

The main recommendation is that CLARIN and CESSDA look together towards offering a mixed solution also investigating other possible options than they currently rely on and support.

Note (added in v1.2). In Q1 2021 CLARIN decided to adopt Skosmos as a vocabulary platform which strengthens considerably the position of Skosmos as a vocabulary platform standard within the SSH. Motivations for this choice were the proven technology and broad use also within the SSH organisations. SKOSMOS is developed by the National Library of Finland also for its own use, adding to expected lon-term availability. Nevertheless the variety of use-cases still offers room for multiple solutions. A "federated approach", linking the different existing SSH Skosmos instances and the CESSDA platform based on the Skosmos API, that is also supported by other platforms should be further investigated.

4. The CLARIN vocabulary initiative and the SSHOC community involvement

A definitive final SSHOC recommendation in the form of a ready-to-use solution did not emerge from the initial task 3.1 inventory when the analysis carried out in WP3.1 confirmed, once again, that varied practices in how vocabularies are used for enabling access to research content are dominating the SSH. This scenario called for a unification effort and CLARIN has taken the initiative to (i) collect, register, and harmonize SSH vocabularies/terminologies/taxonomies and to (ii) improve the service offer of vocabulary platforms in the framework of SSHOC.

In the framework of the Vocabulary initiative, as part of the SSHOC activities and of WP3.1, CLARIN has organised a series of virtual information sessions where partners of SSHOC and members of the Social Sciences and Humanities community have been involved in order to discuss:

1. leading vocabulary management and publication platforms, i.e. Wikibase, Skosmos, and the CESSDA Vocabulary service, and raise awareness about their use in the SSH community
2. the SSHOC recommendations for Vocabulary Platforms (presented in this report).

The first set of SSHOC virtual sessions provided an update about the different vocabulary platforms and the vocabulary-related activities, and discussed the potential overlapping. The sessions took place during September 2020 and the results of the consultations are available at the SSHOC website[5].

A final virtual workshop was organized on 6 November 2020 to discuss the SSHOC Considerations for Vocabulary Platforms. The event followed up on the series of webinars, raising further awareness not only about the harmonisation of the vocabulary-related activities within SSHOC but also about the importance of maximizing access to research contents and to improve discovery in Social Sciences and Humanities.

The organizers (Van der Lek, Monachini, Van Uytvanck, Broeder, Fišer from CLARIN) set up this workshop as a part of WP3.1 and WP6.2 activities. The event started with the introduction about the background of the vocabulary initiative within SSHOC, followed by the presentation of the first outcomes of the MS08 Report. The introductory presentations were followed by an overview of the SSHOC WPs updates on vocabularies, after which CESSDA, CLARIN and DARIAH shared their experience with controlled vocabularies (CVs) and vocabulary platforms. The virtual workshop closed with a panel discussion where several experts from SSHOC and TRIPLE evaluated whether the CVs and vocabulary hosting platforms could be made more interoperable by following the FAIR principles. The results of this consultation are available at the CLARIN website[6].

About 60 participants attended the event: most of them came from the Netherlands (17%), France (12.8%), Italy (11.6%) and Germany (9.3 %). These numbers only account for the on-line participants; the overall reach is higher since it also includes the views of the recordings.

The report of this event forms an integral part of this Milestone Report and is attached here as Appendix 4 – SSHOC Consideration for Vocabulary Platforms.

The initiative, especially the information-sessions and workshop, also provided information and excellent opportunities to consider alternative platforms, as this was relevant for CLARIN.

---

[5] Agenda and slides of Vocabulary Initiative Information Sessions, https://www.sshopencloud.eu/sshoc-online-information-sessions-open-source-vocabulary-hosting-and-management-platforms-first, https://www.sshopencloud.eu/sshoc-online-information-sessions-open-source-vocabulary-hosting-and-management-platforms-second, https://www.sshopencloud.eu/sshoc-online-information-sessions-open-source-vocabulary-hosting-and-management-platforms-third , accessed 30 March 2021.

[6] SSHOC Considerations for Vocabulary Platforms - workshop programme and slides , https://www.clarin.eu/event/2020/sshoc-considerations-vocabulary-platforms-virtual-workshop , accessed 30 March 2021.

References

[1] Brugman H, Lindeman M Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service, http://www.catchplus.nl/wp-content/uploads/2012/11/Brugman-OpenSKOS-final.pdf

[2] Suominen, O., Ylikotila, H., Pessala, S., Lappalainen, M., Frosterus, M., Tuominen, J., Baker, T., Caracciolo, C., Retterath, A. (2015). Publishing SKOS vocabularies with Skosmos, http://skosmos.org/publishing-skos-vocabularies-with-skosmos.pdf

[3] A. Stellato, A. Turbati, M. Fiorelli, T. Lorenzetti, E. Costetchi, C. Laaboudi, W. Van Gemert and J. Keizer Towards VocBench 3: Pushing Collaborative Development of Thesauri and Ontologies Further Beyond http://ceur-ws.org/Vol-1937/paper4.pdf

[4] http://semanticturkey.uniroma2.it

[5] https://www.ics.forth.gr/isl/index_main.php?l=e&c=243

[6] Bandholtz T, Schulte-Coerne T, Glaser R, Fock J, Keller T, iQvoc – Open Source SKOS(XL) Maintenance and Publishing Tool, http://ceur-ws.org/Vol-699/Paper2.pdf

5. Appendix 1 – *Platform evaluation screenshots*

System name and URL	Evaluator	Data model	Notes on evaluation	Concepts	Synonyms of concepts	Hierarchy (broader, narrower, related terms)	Note fields, e.g. definitions or use notes	Concept schemes	Multilinguality and translation	Import & Export	Browser	Editor	User management	Documentation of software	Linking/Mapping/Alignment	Examples of current use	Versioning and history
openSKOS	Austrian Centre for Digital Humanities, Vienna	SKOS; + openSKOS custom data model to extend SKOS	ACDH had important issues with tests, would need to ask them what they were.	Preferred labels and notations must to be unique					Language: English	New version: import via API. Accept only one collection/concept per call. Import follows strict workflow: 1. create concept scheme; 2. create collections; 3. create concepts in order (e.g. if concept1 skos:broader concept2, concept2 must already be imported beforehand).	Simple tablelike view for each concept; main drawback - links are not active (in recent version)	No web editor in new version, but editing possible via PUT requst	Openskos introduces Tenant (Institution)	github: https://github.com/OpenSKOS/OpenSKOS/tree/developMeertens API docs: http://editor.openskos.org/apidoc/index.html#api-Concept-CreateConcept			
openSKOS 2	Meertens Institure/CLARIN	SKOS, optionally SKOS-XL; + openSKOS custom data model (tenants and sets) to extend SKOS		Preferred labels and notations must be unique in a ConceptScheme conform https://www.w3.org/TR/skos-reference/#L2646	supported	supported	supported	supported	supported	import via API or command line	separate browser, or guest access to the editor Comment from Taina: judging from CLARIN concept registry, the browser does not seem to be totally user-friendly - no way to organise concepts alphabetically, search all beginning with a certain letter, or to choose the number of results displayed on a page.	editor with support for roles	tenants (Institutions) and roles	github: https://github.com/OpenSKOS/OpenSKOS API docs: http://editor.openskos.org/apidoc/index.html	New API under development: https://github.com/OpenSKOS/API/blob/develop/doc/OpenSKOS-API.md	https://vocabularies.clarin.eu/clavas/ (uses only the API) http://openskos.beeldengeluid.nl/api/ (only API publically available)	concepts can have states. e.g., candidate and approved; redirected can be used for simple versioning

Image 1a – Platform Evaluation

CESSDA Vocabulary Service	CESSDA/Taina (FSD)	SKOS (but not native SKOS) with additional metadata	One possible use mode: the service can be used as an Editor for suitable vocabularies and then vocs can be harvested into the common publication platform chosen. The vocabularies can also be published in the service. May not be best option for large vocabularies with hundreds of terms. Tool created by CESSDA/GESIS.	Support. Concept/code values must be unique within a vocabulary (concept scheme), not across vocabularies.	No at this point, but may have later.	Supports broader and narrower terms not related terms	Definitions for each concept/code. CV level elements: - Definition of the CV itself - General note on the CV - Usage note, e.g. used for a particular DDI documentation standard element - License - Copyright statement - Agency - Translating agency - Version number - Version changes information - Automatically produced citation - URI of the	Can contain several vocabularies. Grouped by creating agency. Each controlled vocabulary is one ConceptScheme; the vocabulary may contain different language versions but each has the same ConceptScheme as the source vocabulary.	Each language version is translated separately to the common Concept Scheme. CV name, CV definition and note, concept/code descriptive term and concept/code definition can be translated. Usage info copied from source to translations but can be edited (e.g. translated). Each vocabulary and each code have a machine-actionable value (CV Short name, value of the code) that is not translated and stays the same across languages,	Can import concepts/codes in csv (from Excel). Export: vocabulary export available in PDF, html and SKOS, user can choose which languages to include in the export There is also a merged API with all language versions included.	Provides online search UI for general users. Allows searching current versions of published vocabularies by vocabulary name or term. Filtering by agency or by language. Allows exports of published vocabularies. There is another online search for registered users with a role. This allow searching	Online Editor.	Yes, access managed by roles. One person may have more than one role and in more than one agency. Other than super admin roles are agency and language specific. Roles include administrator of source vocabulary, administrator of translated vocabulary, contributors which can amend their agency vocabularies in specified language in certain stages (DRAFT, INITIAL REVIEW).		No	Used as an Editor for DDI Alliance vocabularies which are then published also on the DDI Alliance website, in addition to being available in the CVS. Used for CESSDA Vocabularies but currently there is another software for the CESSDA thesaurus.	The versioning system is based on the assumption that the vocabularies are used in metadata and also in filters in search interfaces, so there needs to be detailed information on what has been changed in a new version. Source vocabularies (SL) have two-digit version number to denote major and minor changes. If there is a major change, first digit changes, if a minor, the second digit changes. E.g. 3.2. Version number is suggested by the system but can be manually edited. Translated vocabularies (TL) have three-digit version number where the two first digits are the source version number and the third a running version number. E.g. 3.2.2 means a second translation version of the source vocabulary 3.2. Only the third digit is manually editable. When a new version of a SL vocabulary is published, the system freezes the previous version and all its translations, they can no longer be amended. The system clones the previous version in all languages as a starting point when a new version is being created. The search UI only contains the latest versions but current versions have links to the previous ones. Version information entered into the system specifies what has been changed from one version to next. There is also a comparison table demonstrating the differences. The merged API with all language versions will have a versioning that indicates when there have been changes in any language version, for harvesting purposes.
Themas	Austrian Centre for Digital Humanities, Vienna	THEMA XML	Old-fashion look but functions are quite good. Downside is only one thesaurus at a time (chose a thsaurus duing log-in). So it is best to use it as a backend editor, and SKOSMOS may be used for browsing.				basic functionalities for: - terms; - hierarchies; - facets; - thesauri; - database; - users	Can have multiple vocabularies/thesauri, but work only one at a time (log-in)	Translation is possible for a term and scope note, but there seems to be a bug to use this function	Seems to be possible to import only THERMA XML. Export can be THEMAS XML or SKOS RDF	Rather old-fashioned look&feel, but functions are good	Most important functions are available. It is very easy to move/connect nodes and sub-nodes.	5 roles (Reader, Libraries, ThesaurusTeam, ThesaurusCommittee, and Admin. A bit too much, it would be nice if we can configure and simplify) which correspond to the permissions of change 4 status (ForInsertion, UnderConstruction, ForApproval, Approved)	Not much documentation yet added but here is some https://vocabseditor.acdh-dev.oeaw.ac.at/about/.	basic BT, NT etc support, probably not more complicated ones		
TemaTres	CESSDA	May not allow harvesting the whole vocabulary in SKOS, only by terms.	Did not seem to have a steady development team. Too risky to choose as an Editor, if based on individual efforts.															
iQvoc	CESSDA/GESIS		GESIS has used this for its own thesaurus. They gave up because very cumbersome with language versions. Their advice: forget															

Image 1b – Platform Evaluation

SKOSMOS	Austrian Centre for Digital Humanities, Finnish Social Science Data Archive	SKOS; Dublin core; + SKOSmos custom to extend SKOS; +SKOSXL	No editor (?) Only a browser and publication tool. Good documentation, steady development team, quick to install.	support				vocabulary can contain multiple concept scheme instances	any language; languages have to be added in vocabularies.ttl in vocabulary metadata (e.g. skosmos:language "de","en";)	Import is done in two steps: 1. Upload your dataset to Jena Fuseki (formats accepted: Turtle, RDF/XML or TRiG) 2. Configure vocabularies.ttl on server: add metadata about your dataset (list of required properties https://github.com/NatLibFi/Skosmos/wiki/Vocabularies#vocabulary-configuration-vocabulariesttl) Export: available for a whole vocabulary as RDF/XML, TURTLE (one have to specify links to data dumps in <void:dataDump> in vocabularies.ttl); for each concept as RDF/XML, TURTLE, JSON	autocomplete search, indexes preflabels, altlabels and hidden labels	No Editor (?), is only a browser and publishing tool.	no user management needed	github: https://github.com/NatLibFi/Skosmos Issues tracker: https://github.com/NatLibFi/Skosmos/issues docs: https://github.com/NatLibFi/Skosmos/wiki/Vocabularies	linking among vocabularies is possible; linking among one vocabulary too with skos:related, skos:closeMatch, skos:exactMatch	Unesco thesaurus http://vocabularies.unesco.org/browser/thesaurus/en/z Finnish thesaurus and ontology service http://finto.fi/yso/en/page/p11219 AGROVOC Multilingual thesaurus http://aims.fao.org/standards/agrovoc/functionalities/search - publication on skosmos, Editor is VocBench3	
VocBench	UK Data Service (Darren Bell)	OWL, SKOS, SKOSXL, configuaeble support other classes and attributes as required	Very easy to get running locally on windows. Server side installation using Docker and GraphDB straightforward	Native SKOS Concept class	Encompassed byn native SKOS model (altLabel)	Encompassed byn native SKOS Concept class attributes (not just broader but predicates like broaderTransitive for example)	Native SKOS Concept class attributes.All note types Currently on Version 5 at Feb 2019	Multiple schemes and skos:collections	Supports around two hundred languages on all properties	import and export via GUI in multiple serialisations. no command line needed but can also be done with full native RDF API (RDF4J)	Tree view, but also sparql endpoint implemented with http://about.vasgui.org/	integrated in browser, not very nice looking but seemed to be functional No-code custom forms can be created in v5 and stylesheet easy to alter Google user group: https://groups.google.com/forum/#!forum/vocbench-user/join	many roles, project based	http://vocbench.uniroma2.it/doc/	(maybe) only from locally stored/imported concepts documentation says YES for alighnment support and alignment validation. In Eurovoc, for example, alighments with specialized thesauri Agrovoc, Gement etc.	EU's multilingual thesaurus Eurovoc and AGROVOC use VocBench3 as the editor	Yes, users can create snapshots of a repository and tag them with a version identifier and other metadata. It is possible to time-travel through different versions of the edited repository.
Vocabulary Service		Strictly SKOS model Some of the Dublin core properties for metadata	Created by Austrian Centre for Digital Humanities (ACDH) https://www.oeaw.ac.at/acdh/	Support	?		note scopeNote definition changeNote editorialNote historyNote example	One controlled vocabulary is one Concept Scheme; concepts can be grouped inside Concept Scheme in Collections.					Allows sharing a Concept Scheme with other users so that can create and edit it (all content i.e. concepts and collections).	The software is open source, there is a github repo: https://github.com/acdh-oeaw/vocabseditor Not much documentation added yet but here is some https://vocabseditor.acdh-dev.oeaw.ac.at/about/			All changes are stored in JSON format in the database, the objects can be reverted to previous states in the admin interface.

Image 1c – Platform Evaluation

6. Appendix 2 – *Questions about Vocabularies*

Vocabulary name	TaDiRAH	BackBone Thesaurus (BBT)	DDI Alliance and CESSDA vocabularies	CLARIN Concept Registry (CCR)
Link to the vocabulary	http://tadirah.dariah.eu/vocab/index.php	https://vocabs.dariah.eu/backbone_thesaurus/en/	https://vocabularies.cessda.eu/#!discover	https://www.clarin.eu/ccr
Does the vocabulary need a) an Editor platform for maintaining it, b) a browsing and publication platform to provide access to it, or is the current enough?	Editing is done on Github: https://github.com/dhtaxonomy/TaDiRAH/		The vocabularies are maintained and published in CESSDA Vocabulary Service (link above). DDI publishes their vocabularies on their own website as well. So CESSDA does not need for new vocabulary/ontology platforms. All vocabularies have the same metadata model.	Ideally both, but right now data input is done in batch mode based on tabular data imports
Name of organisation maintaining it	DARIAH external users		CESSDA	CLARIN
Name of the person filling in the information	Yoann Moranville		Taina Jääskeläinen	Dieter Van Uytvanck
What is the vocabulary used for?	For research activities, research objects and research techniques in DH		Used in metadata for describing data, and in CESSDA data catalogues to allow filtering results.	Metadata disambiguation
Who uses it?	openmethods.dariah.eu registries.dariah-dariah.eu jeresah.dariah.eu dirtdirectory.org and more (see github page)		CESSDA Data Catalogue and CESSDA Service Providers.	vlo.clarin.eu
Who can browse/see it?	Anyone		Anyone	Anyone
What is important in browsing, how would the user need to see the vocabulary, i.e. with all language versions of a term together, or each language separately as a whole?	Can't understand the question, sorry		At present, each language version of the vocabulary presented separately.	all language versions of a term together
Approximate number of terms/concepts	121 terms (http://tadirah.dariah.eu/vocab/sobre.php)		From a few terms up to 100, depending on vocabulary.	3163
Who can edit the vocabulary, i.e. all interested parties or should editing rights be restricted to certain persons/roles?	Rights are restricted to maintainers		Only people with a relevant role can log in and access the Editor. Anyone can browse published vocabularies, logged-in users see also the draft versions.	restricted to editor role
Does it have hierarchy, that is, broader and narrower terms? (Yes/No)	Yes		Yes	yes
Does it have related terms? (Yes/No)	No		No	do not know
Do terms have synonyms/alternative labels?	No		No	yes
Does it have definitions?	Partly: Activities are, Techniques and Objects are not		Yes	yes

Image 2a – Questions about Vocabularies

Does the vocabulary exist in different languages/Are terms translated into different languages? Please specify how many languages approximately.	Yes, 4: English, French, German, Spanish		Yes, presently about 7 language versions, more coming in.	I believe so. If yes, nr of languages depends on subcollection (many translations stem from the earlier ISOcat incarnation)
If translated, specify what content/elements are translated (terms, definitions, etc.)	Only terms		CV name, CV definition, descriptive term of the code, definition of the code.	definitions
Who does the translation?	Various people involved in DH and wanting to use the TaDiRAH themselves		CESSDA SPs and non-CESSDA organisations.	CLARIN/ISOcat community
Is import of current vocabularies needed? If yes, what is the format of the vocabulary at present. (This question refers to vocabularies that are now stored in some system for the time being but would in future be maintained in the platform selected in SSHOC)	don't know - but I would say no			I do not understand the question
What export formats are needed?	don't know - but it is available in various export formats (incl. DC, Skos, JSON-LD)		SKOS/rdf, pdf and html.	SKOS
Is API needed?	don't know - for what aim?		yes	yes
Is versioning needed, i.e. does the vocabulary have version number and publication date, or is it updated continuously?	Yes, it was updated, but not recently		Strict versioning, source language has two-digit version number (major and minor changes,e.g. 2.2), each translated version has its own three-digit version number but the first two digits are the source version number (e.g. 2.2.1).	Not sure (Taina: does CCR have any versioning at concept level, that is, does the user know when a change to the concept-related information has been made?)
Do users need to know what has been changed from last publication?	Not need, but should		Full documentation of changes.	Not sure
Does the vocabulary have a license?	don't know		CC-BY 4.0	CC-BY

Image 2b – Questions about Vocabularies

7. Appendix 3 – *Software Platform Evaluation screenshot*

System name and URL	Software as a Service (SaaS)	Standalone installation	Interoperability Protocols	Linked Data	API	Source Code	License
openSKOS	Yes	Yes	OAI-PMH	API	http://openskos.org/api	https://github.com/OpenSKOS/OpenSKOS	GNU GPL 3.0
SKOSMOS	No	Yes		API/SPARQL	http://api.finto.fi/doc/#/	https://github.com/NatLibFi/Skosmos	MIT License
VocBench	No	Yes	RDF4J	SPARQL		https://bitbucket.org/art-uniroma2/vocbench3/src/master/	
ACDH Vocabularies Editor	No	Yes				https://github.com/acdh-oeaw/vocabseditor	MIT License
CESSDA Vocabulary Service	Yes	No				https://github.com/isl/THEMAS/	EUPL 1.1
Themas	No	Yes					EUPL 1.1
TemaTres	No	Yes		SPARQL	https://r020.com.ar/tematres/demo/api/		GNU GPL 2.0
iQvoc	No	Yes		SPARQL?		https://github.com/innoq/iqvoc	Apache License, Version 2.0

Image 3 – Software Platform Evaluation

8. Appendix 4 – *SSHOC Considerations for Vocabulary Platforms*

Appendix 4 is a contextualized compilation from different sources, but mainly originating from the workshop notes[7], and therefore has a different style of citing and linking to external sources which the team prefers to present in this form for reference.

Background

On November 6, CLARIN organised a virtual workshop to discuss the *SSHOC Considerations for Vocabulary Platforms* as part of the SSHOC Tasks 3.1 and 6.2.

The event wrapped up a series of webinars that CLARIN started to organise in September as part of their initiative to collect, register, and harmonise SSH controlled vocabularies, thesauri and taxonomies, and to improve the service offer of vocabulary platforms. This initiative emerged from the need to harmonise not only the vocabulary-related activities within the SSHOC work packages but also to unify the access to research contents and improve discovery in Social Sciences and Humanities.

Session Overview & Format

The virtual workshop, streamed via ZOOM, consisted of a total of eleven presentations (10-20 min long) and a discussion panel. Each session concluded with a lively Q&A part.

The full programme overview is available on the CLARIN website and includes the presentations and recordings.

About the organisers

CLARIN set up the virtual workshop as part of their tasks in SSHOC (T3.1 and T6.2).

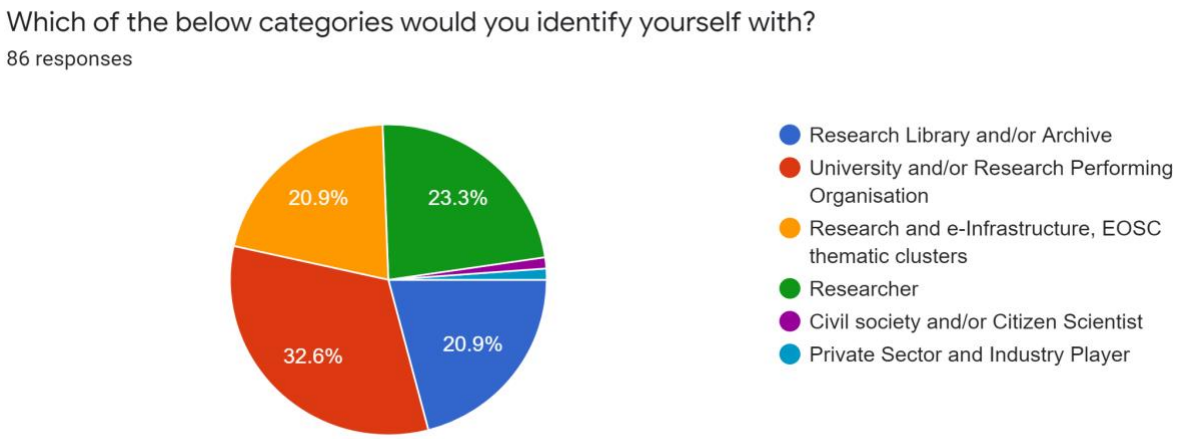
- Iulianna van der Lek, Project Coordinator of the Vocabulary Initiative at CLARIN ERIC
- Monica Monachini, CLARIN - IT National Coordinator and SSHOC T3.1.
- Dieter Van Uytvanck, Technical Director at CLARIN ERIC
- Daan Broeder, Project Manager at CLARIN ERIC
- Darja Fišer, CLARIN/UL-FF, SSHOC T6.2

[7] Iuliann van der Leck, Kristina Pahor de Maiti. Workshop notes: SSHOC requirements – Vocabularies and Vocabulary Management, https://www.sshopencloud.eu/news/workshop-notes-sshoc-requirements-vocabularies-and-vocabulary-management-platforms [accessed 11.06.2021]

Participants

A total of 86 people signed up for the event, out of which about 60 participants attended. Most participants came from the Netherlands (17%), France (12.8%), Italy (11.6%) and Germany (9.3 %). It should be noted, however, that these numbers only account for the live viewers, and that the overall reach is bigger since it also includes the views of the recordings.

About 32.6% of the total registered participants came from universities and research-performing entities, about 23.3% were researchers, and 20.9% worked for research and e-Infrastructure clusters or research libraries and archives. Only 1% of the participants represented the private sector.



Event summary

One of the main goals of SSHOC Task 3.1 *Multilingual Terminology*, as described in the DoA of WP3, is to find a suitable vocabulary server and publication platform to improve accessibility and discovery by non-native speakers. Therefore, the virtual workshop started with a presentation of the MS08 Report results: *Choice of vocabulary publication platforms for SSHOC*. The presentation was followed by an overview of the SSHOC WPs updates on vocabularies, after which CESSDA, CLARIN and DARIAH shared their experience with controlled vocabularies (CVs) and vocabulary platforms. The virtual workshop closed with a panel discussion where several experts from SSHOC and TRIPLE evaluated whether the CVs and vocabulary hosting platforms could be made more interoperable by following the FAIR principles.

Presentations & discussions: Key points

The following section outlines the main points of each presentation and the panel discussion.

Prof. Dr. Franciska de Jong, executive director at CLARIN ERIC, opened the workshop explaining the background of the vocabulary initiative. This initiative has emerged from the need to align the vocabulary activities across the SSHOC work packages, and it will help optimise the sharing of research data across various practices and domains.

Prof. de Jong further presented the work plan of the vocabulary initiative. In the early stages of the SSHOC project, WP3 collected the SSHOC requirements for the vocabulary registry in the milestone report MS8.0 *SSHOC considerations for vocabulary platforms*. In September, CLARIN launched a series of virtual events to raise awareness in the Social Sciences and Humanities (SSH) community about vocabulary hosting and publication platforms. Experts from SSHOC and other related Horizon-2020 projects, as well as end users, presented use cases and exchanged best practices. Future activities will consist of a re-evaluation of the first milestone, an inventory of the most known controlled vocabularies/taxonomies, and potential matching of vocabularies and their metadata.

Link to the recording:

Introduction (recording)	**Franciska de Jong** Executive Director CLARIN ERIC

WP3.1 MS8 report: Choice of vocabulary publication platforms for SSHOC

Monica Monachini, SSHOC WP3.1 leader and the CLARIN-IT National Coordinator, presented the results of the MS8.0 report: *Recommendations for Vocabulary platforms in SSHOC*. The main aim of WP3 is to lift technologies and services in the SSH Cloud by contributing to infrastructure components and content, and making the CLARIN technologies useful for the other SSH infrastructures. Within this context, the goal of WP3.1 is to provide resources and tools to improve discovery of the SSH research data and facilitate reusability. The team will translate the metadata in different languages and find a suitable vocabulary publication platform.

The milestone report MS8.0 provides a set of recommendations for vocabulary publication platforms for SSHOC. Through focused surveys, interviews with experts, the task team evaluated several platforms and produced a set of requirements that a vocabulary service should have, namely:

- Import and export of thesauri in SKOS / RDF format;
- Unified access to all vocabularies;
- Editing with collaborative functionalities;
- Alignment functions between vocabularies and external resource;
- Terminology management interface (hierarchical structure, semantic relationships translations, facets);
- Management of different roles and workflows;
- Management of versioning;
- API services;
- Friendly and intuitive interface, suitable for use by non-expert users;
- Flexible to be adapted to new needs and standards.

Three out of the eight platforms that have been evaluated seem to be the best candidates to host and publish the SSHOC vocabularies, namely ACDH-CH, CESSDA and CLARIN Vocabulary Services. However, none of them seem to fulfil all the requirements. Hence, more investigations are needed. The insights

collected during the virtual sessions and the workshop will help refine the requirements and produce suitable recommendations.

The participants were interested to learn how WP3 envisions the collaboration between the SSHOC platform for CVs and the existing providers of vocabulary platforms.

Links to the presentation slides and recording:

Presentation of the MS08 results: *Choice of Vocabulary publication platforms for SSHOC (*slides) (recording)	**Monica Monachini** CLARIN - IT National Coordinator and SSHOC WP3

SSHOC updates on vocabularies

In this one-hour session, five SSHOC WPs gave a short update about their work and experience with semantic artefacts, such as controlled vocabularies and taxonomies. Since some of the SSHOC WPs did not include specific controlled vocabulary tasks, the presenters shared their experience acquired in other Horizon 2020 projects, such as ARIADNE and DARIAH.

SSH VOCABULARY SURVEY

In a joined presentation, Clara Petitfils & Nicolas Larrousse (WP7-WP3) presented the results of the SSH vocabulary survey, and the resources needed for SSH Open Marketplace content description.

The survey was open from February until May 2020 and contained 16 questions about the use of SSH vocabularies, alignment, languages, availability and maintenance. The results revealed the semantic artefacts that are often used by the community: the Data Documentation Initiative (DDI), Getty Art & Architecture Thesaurus (AAT), CESSDA Controlled Vocabularies, ELLST, and Dublin Core. The respondents wished the vocabularies they used were matched with Getty and Wikidata.

Nicolas Larousse pointed out that vocabularies are essential for the SSH Open Marketplace to describe the entries, improve search and retrieval, and foster discoverability. Furthermore, the external sources ingested into the marketplace will need to be aligned with the "local" vocabularies.

VOCABULARY MAPPING TOOL FOR ARCHAEOLOGY IN ARIADNE PLUS

Holly Wright (WP5) from the Archaeology Data Service presented how they achieved vocabulary mapping in ARIADNE, another Horizon-2020 project. With 1.9 million data records aggregated, the research project produced a total of 6 416 vocabulary mappings when aligning subject vocabularies to Getty AAT.

The vocabulary mappings in ARIADNE cut out the linguistic barriers to cross-searching, such as language, spelling, homonyms, synonyms and level of specificity. Wright reiterated that it is important that the source datasets are produced with aggregation, cross-search and reuse in mind. She concluded that

although there is a strong need to harmonise meaning, mapping everything to everything is nearly impossible.

To be able to match local subject terms and concepts to Getty AAT concepts, the research team developed a vocabulary matching tool in the first phase of the ARIADNE project.

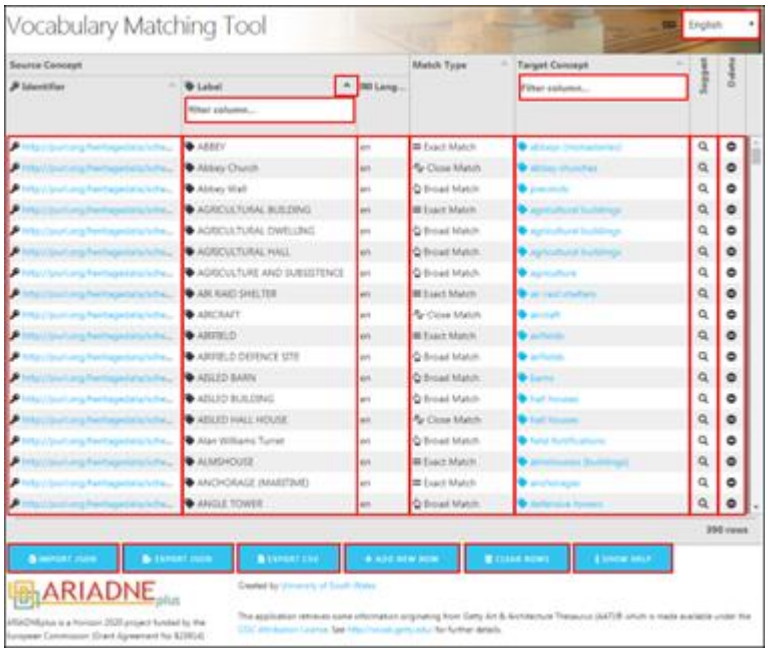


Image source: Vocabulary Matching Tool - Help (d4science.org)

The tool supports the SKOS mapping properties, and it has a multilingual user interface. Researchers can use it to search and browse through AAT vocabularies to make more informed decisions about the mappings. Users can export their data to JSON or delimited text (CSV) format that they can import in other applications. If users encounter terms/ concepts without URIs, they can add them manually to facilitate Linked Open Data source vocabularies.

MANAGEMENT OF CONTROLLED VOCABULARIES WITHIN THE AIOLI PLATFORM WITH THE OPENTHESO TOOL

Isabelle Cao from CNRS-MAP presented how they create and manage controlled vocabularies in SSHOC WP4.6 within the Aioli platform using the Opentheso tool.

Aioli is a reality-based 3D annotation platform for the collaborative documentation of cultural heritage. It supports SKOS/RDF, and it is useful to reconstruct images automatically in 3D and annotate them by adding both custom and computer descriptors. The platform integrates several controlled vocabularies, which helps users consistently describe annotations.

The platform uses Opentheso, an ISO-compliant web-based thesaurus management tool, to create and manage controlled vocabularies. Data can be imported and exported in standard formats. The concept management interface is intuitive and supports hierarchical relationships, synonyms and translation. The thesaurus navigation allows for different views: tree, hierarchical, collection and index. Users can

align concepts with other thesauri using semi-automatic alignment and pre-configured sources, such as Wikidata, Geonames etc. It is possible to add candidate concepts collaboratively, add notes and translate them. The system automatically adds the validated terms to the thesaurus. Finally, yet importantly, the platform offers features for user management and collaboration.

Opentheso has been implemented in the Aioli platform using a docker container. Users can manage thesauri directly in the Aioli platform using the Lexicon function. The published thesauri automatically synchronises with Opentheso. Future developments will include a search engine that will allow users to search for annotations within the project itself. Finally, there are plans to make the UI and the documentation available in English as well.

To learn more about the technical specifications of the Aioli platform, see deliverable D4.16 Specification of the new feature of the Aioli platform.

BackBone Thesaurus, a model for thesauri interoperability

Because SSHOC T4.7 does not include tasks directly related to controlled vocabularies but ontologies, Chryssoula Bekiari and Eleni Tsoulouha (WP4.7) shared their experience with the BackBone Thesaurus (BBT), a platform developed under Horizon-2020 project, DARIAH EU.

According to the BBT Maintenance WG, the platform aims "to establish a coherent overarching thesaurus for the humanities, under which specialist thesauri and structured vocabularies used across scholarly communities can be aligned and form a thesaurus federation.'' The aligned vocabularies keep their autonomy, and the users can access them through global thesaurus. BBT follows the CIDOC-CRM classification, and users can access the federated thesauri through the ACDH vocabulary repository service.

The platform supports high-level concepts, facets and hierarchies that do not exhaust the domain they classify. The BBT platform integrates BBTalk, a multilingual online service for thesauri management and maintenance. The service developed by FORTH-ICS supports RDF, and it includes a thesaurus alignment tool, collaborative features and keeps track of versioning.

How Dataverse supports external vocabularies

Slava Tykhonov (WP5) from the Data Archiving and Networked Services (DANS) in the Netherlands presented how Dataverse supports controlled vocabularies with focus on semantic and technical interoperability, as defined by EOSC Interoperability Framework v1.0. Dataverse is a product developed by Harvard University and used in SSHOC T5.2. A recent self-assessment analysis performed by Merce Crosas[8] of the implementation of the FAIR principles in Dataverse has revealed that while scoring high

[8] Merce Crosas, 2020, "FAIR principles and implementation in Dataverse": FAIR-Dataverse-Tromso (harvard.edu) [acessed 30 Mar 2021]

on Findable, Accessible and Reusable, the product is weak from an interoperability point of view. To tackle this, Dataverse aims to implement a FAIR metadata schema and connect the metadata to ontologies and controlled vocabularies.

Using GRID (Global Research Identifier Database) in SKOS, the platform provides a convenient depositor web interface to link the metadata of the datasets stored in the Dataverse network to external controlled vocabularies. The development team also plans to investigate if it is possible to disambiguate concepts using NLP tools and create the links between Dataverse and external vocabularies automatically.

The platform uses the Skosmos API specification protocol to ensure technical interoperability with other controlled vocabulary services (CESSDA). It connects to a Semantic Gateway application that enables users to query the vocabularies stored on different platforms (e.g. Skosmos). Users can link every field in Dataverse to the appropriate vocabulary following the FAIR principles. Finally, the platform supports multilinguality, and it allows researchers not only to enrich metadata but also export it to Linked Open Data Cloud to increase its findability or use it to train other Machine Learning models.

Links to the presentation slides and recordings of the SSHOC WPs updates on vocabularies session:

SSHOC WPs updates on vocabularies: • Vocabulary Survey (recording) • Vocabulary mapping tool for archaeology in ARIADNE plus (recording) • Management of controlled vocabularies within the Aioli platform with the Opentheso tool (recording) • BackBone Thesaurus, a model for thesauri interoperability (recording) • How Dataverse supports external vocabularies (recording) (slides)	**Clara Petitfils & Nicolas Larrousse (WP7-WP3)** **Holly Wright (WP5)** **Isabelle Cao (WP4)** **Chryssoula Bekiari (WP4)** **Slava Tykhonov, (WP5)**

Overview of the CLARIN & SSHOC webinar series

Iulianna van der Lek, the CLARIN coordinator of the vocabulary initiative, gave an overview of the three vocabulary information sessions that took place in September: Introduction to Wikibase, CESSDA Vocabulary Service and Skosmos. A separate CLARIN-SSHOC report of these events summarises the sessions.

CESSDA: main requirements and best practices

Taina Jääskeläinen from the Finnish Social Science Data Archive gave an overview of the vocabularies available in the CESSDA Vocabulary Service, the platform requirements and the tools they currently use.

The following controlled vocabularies are available within the service: subject-specific vocabularies, such as Thesaurus ELSST and the CESSDA Topic Classification, vocabularies of research methods created by DDI Alliance, and vocabularies used for the CESSDA Data Catalogue.

The vocabulary service platform supports advanced user management, vocabulary management, translation and publication, concept hierarchy and synonyms. A new requirement concerns the CESSDA organisations producing metadata from varied disciplines that would benefit from having concepts specific to a research domain within a broader controlled vocabulary. This new requirement needs to be further specified.

Links to the slides and the recording:

CESSDA: main requirements and best practices (slides) (recording)	**Taina Jääskeläinen** Finnish Social Science Data Archive

CLARIN: main requirements and best practices

CLARIN has been managing vocabularies with CLAVAS vocabulary server since 2017. Currently, they are looking for a follow-up vocabulary management and publication platform. The requirements for the new platform are the following: it should be a sustainable open-source solution, SKOS compliant, with a GUI and a vocabulary editor, advanced browsing functionalities and fast look-up via API, and it should support persistent identifiers and multilinguality. Preferably, the solution should have preloaded and curated controlled vocabularies, such as ISO-693-3.

Different vocabulary platforms could be aligned through a compatible API, multidirectional synchronisation and double curation work.

The interoperability challenges between vocabularies could be tackled by reusing existing curated vocabularies and linking to authority files wherever possible. This approach would avoid reinventing the wheel over and over again.

Links to the presentation slides and the recording:

CLARIN: main requirements and best practices (slides) (recording)	**Dieter Van Uytvanck** Technical Director at CLARIN ERIC

DARIAH: main requirements and best practices

Matej Ďurčo (ACDH-CH) and Laure Barbot (SSHOC) shared the requirements and best practices for vocabularies and vocabulary platforms that they learnt from their DARIAH activities and how these could

benefit the SSH Open Marketplace. DARIAH set up the WG Thesaurus Maintenance (BackBone Thesaurus), contributed to the vocabulary activities in PARTHENOS, and have their central vocabularies service.

According to M. Ďurčo, there should be a clear distinction between requirements for vocabularies and requirements for vocabulary hosting platforms.

Vocabularies should be published as Linked Open Data (based on SKOS data model) and provide comprehensive coverage of the domain through concept definitions and examples. Users should be able to reuse existing vocabularies or link them to other artefacts, thus ensuring semantic interoperability. Finally, vocabularies are useless if they are not available in authoring environments.

The vocabulary platforms, on the other hand, should support at least full SKOS data model, implement a curation workflow with full provenance and an API for easy access and look-up.

Matej Ďurčo shared best practices based on TaDiRAH (Taxonomy of Digital Research Activities in the Humanities) and the SSHOC Marketplace. The TaDiRAH vocabularies are available in various applications such as SSK, SSHOC Marketplace and DH Course Registry. The presenter revealed that it was challenging to integrate those vocabularies which did not have a URI and were not available via API.

In SSHOC, DARIAH has the task to develop the SSH Marketplace in WP7 that aims to be a discovery platform for tools, services and resources useful for research activities in Social Sciences and Humanities. Matej reiterated the controlled vocabularies are needed to describe and classify the items available via the platform and facilitate retrieval, browsing and interoperability. He mentioned that SSHOC D7.1 provides the first list of controlled vocabularies that could be part of the system specification of the marketplace, e.g. IANA mime, TaDiRAH, CESSDA Topic Classification, ISO 639-1, etc.

Custom-based properties in the data model and specific workflows for vocabulary creation and curation are under development. PoolParty, a commercial taxonomy and vocabulary management server, is currently used for the hosting, management, and mapping of vocabularies. It seems that candidate concepts pose some challenges because of the communication between the ingest script, marketplace core and the Vocabulary Manager.

The presentation concluded with a snapshot of the vocabulary management workflow in ACDH-CH vocabulary service that uses Skomos to store and publish the vocabularies. Another challenge that we need to address is the following: What do we do when users want to manage the vocabularies directly within their applications, while they are stored and edited in third-party client applications?

Finally, yet importantly, Matej highlighted that it is essential to share the knowledge about vocabularies via dedicated training material, e.g. the Controlled Vocabularies and SKOS e-learning course available on DARIAH-CAMPUS.

Links to the presentation slides and the recording:

DARIAH: main requirements and best practices (slides) (recording)	**Matej Ďurčo** ACDH-CH

Panel: Improving the FAIRness of SSH Vocabularies

The theme of the panel was: "How can we make the SSH vocabularies FAIRer." The theme is in line with the SSHOC goal of helping SSH researchers integrate their work and results according to the FAIR guiding principles scientific data management and stewardship.

Daan Broeder, project manager at CLARIN ERIC, moderated the panel of discussion. The panellists consisted of the following experts: Menzo Windhouwer (KNAW), Suzanne Dumouchel (EOSC), Matej Ďurčo (ACDH-CH), Melanie Bunel (HumaNum).

The discussion revolved around the fact whether the FAIR principles could be applied to semantic artefacts as well. The moderator asked the panellists to reflect on the following topics:

- Would selecting and sharing a single vocabulary platform make the vocabularies FAIRer?
- What measures can further improve FAIRness of the SSH vocabularies?
- How do the editorial/curation processes influence vocabulary FAIRness?

Would selecting and sharing a single vocabulary platform make the vocabularies FAIRer?

Sharing a vocabulary platform can mean either sharing a single instance or sharing the code and creating federated instances.

While using a single platform could help overcome some of the current challenges that vocabulary platforms have, such as the lack of a uniform API to access vocabularies, the panellists agreed that it is something that it is not feasible to achieve. According to Matej Ďurčo, it would be challenging to unify the community and store all the vocabularies on one platform because there will always be different players, stakeholders, researchers who will prefer to be in control of their own platforms and instances.

Suzanne Dumouchel agreed with Matej stating that there should be at least a common place where all the existing SSH vocabularies are gathered and made more visible. The moderator suggested that we could share at least the same publication platform, but highlighted that the curation processes are quite different. He gave Skomos as an example.

Menzo Windhouwer proposed that if the SSH communities are not able to share the same vocabulary platform, we could at least try to reach a consensus on a small common API to be supported by the endpoints, for example, autocomplete to drill down into a vocabulary and getting information on a specific vocabulary item. Such a shared minimal API could help tackle interoperability challenges.

What measures can further improve FAIRness of the SSH vocabularies?

Findability

The first FAIR principle is to make data findable. The moderator encouraged the panellists to reflect on the findability aspects of vocabularies. Besides assigning a globally unique and persistent identifier, it is also important to describe the vocabularies in a coherent way. It does not mean that they need to be published in one instance because there are always ways to harvest the metadata and create a shared catalogue of all the available vocabularies to discover them.

Matej Durco pointed out that researchers prefer to develop their vocabularies, but without following the current standards and guidelines for vocabulary development. The vocabulary service and the editing interface should guide the researchers in their work and support them to develop vocabularies in a consistent and standardised way. This will foster findability and reusability.

Accessibility

The second FAIR principle concerns data accessibility. This can be achieved via authentication and authorisation procedures.

Suzanne Dumouchel suggested that one way to make the vocabularies FAIRer could be by replacing the A for *Accessible* by A for *Adaptable* because we need to keep up with the changes and the evolving vocabularies. She also underlined the need for continuous cooperation with the research community in order to include new concepts., as well as to identify and define the SSH disciplines properly.

Interoperability

The third FAIR principle states that data need to interoperate with other applications or workflows for analysis, storage, and processing[9].

In the context of vocabularies, the moderator pointed out that there are tools to make the content of the vocabularies interoperable and support the mapping process between the vocabulary terms/concepts. It may be worthwhile to take a look at these tools when dealing with large vocabularies in the SSH community.

According to Suzanne Doumuchel, the SSH community could benefit from a project dedicated to SSH vocabularies and that CLARIN could play a role in it. Such a project would help all stakeholders reach decisions and agreements on SSH vocabularies (generic vocabularies and domain-specific vocabularies) and make them more interoperable (for example, using more Wikidata).

[9] FAIR Principles - GO FAIR (go-fair.org) [accessed 11.06.2021]

Matej reiterated that it is essential to distinguish between semantic and technical interoperability. Semantic interoperability could be achieved through an approach that counts for plurality, for example, letting researchers create their vocabularies and then use tools to link to match them.

Since it is time-consuming to map entire vocabularies, the moderator argued in favour of a pragmatic approach; for example, mapping only those parts of vocabularies that are relevant. He referred the audience to the SEMAF report[10] that proposes a flexible semantic mapping framework targeted at specific interoperability goals.

Menzo Windhouwer pointed out that the end users do not know what types of relationships knowledge engineers apply during vocabulary mappings. He proposed to set up a quality assessment of the mapping process that includes some provenance metadata with confidence metrics.

One member of the audience, Andrea Scharnhorst, remarked that discussions around semantic interoperability have been going on for a long time and that there are similar projects taking place in parallel. Hence, we should avoid reinventing the wheel and encourage the SSHOC researchers to build on existing initiatives as much as possible. She gave as examples, the Linked Open Vocabularies (LOVs), the BARTOC Vocabularies, and the KOS Observatory for Social Sciences and Humanities, a project developed by DANS. The latter includes about 125 KOSs from the Social Sciences and Humanities (SSH), and the Life Sciences (LS) that researchers have mapped by applying a thorough methodology.

Sustainability

Menzo Windhouwer asked how the vocabularies could be maintained over time, for example, after the funding received for a research project had ended. He pointed out that it would be laborious to keep the resources up to date by using volunteers because they need specific knowledge engineering skills.

Franciska de Jong, the executive director of CLARIN ERIC, confirmed that sustainability is indeed an issue. She indicated that this topic is a recurrent theme at several layers of SSHOC as a project. Some parties may have the possibility to sustain the tools and linguistic resources that they are developing. For example, CNR has committed to hosting the ARIADNE mapping tool for at least five years after the SSHOC phase has finished. There are plans to make all the mapped vocabs available as RDF downloads for reuse as well. BBT and BBTalk will continue to be maintained voluntarily by the Thesaurus Maintenance Working Group.

Franciska hopes that this vocabulary initiative will motivate the SSHOC WP leaders to work towards interoperability. SSHOC could develop a couple of use cases to demonstrate how WPs have achieved interoperability across platforms.

[10] SEMAF end report now published http://doi.org/10.5281/zenodo.4651421 [accessed 30 Mar 2021]

How do the editorial/curation processes influence vocabulary FAIRness?

Menzo Windhouwer pointed out that the end users do not know what types of relationships knowledge engineers apply during vocabulary mappings, so he proposed to set up a quality assessment of the mapping process that includes some provenance metadata with confidence metrics.

The panel concluded that more discussions are needed to come to a final recommendation for vocabulary platform(s) in SSHOC. Most likely, there will not be a single shared platform to manage the vocabularies. The moderator encouraged the SSHOC WPs to think about how they can achieve interoperability between the different platforms, for example, data exchange, vocabulary identification schemes, vocabulary maintenance and quality management.

Link to the recording:

Panel: SSHOC Considerations for the Vocabulary platforms (recording)	**Moderator**: Daan Broeder **Panelists:** Menzo Windhouwer, Suzanne Dumouchel, Matej Ďurčo, Melanie Bunel

Outcomes & feedback

Outcomes

While the vocabulary virtual sessions we organized in September were meant to be informative, the workshop provided a more in-depth discussion about vocabularies and vocabulary platforms to be used in SSHOC. One recurrent theme was that the SSH community should avoid reinventing the wheel and build on existing initiatives as much as possible. There should also be a clear distinction made between the requirements needed for vocabularies and requirements needed for vocabulary platforms.

REQUIREMENTS FOR VOCABULARIES

Semantic and technical interoperability seem to be the most important requirements. This could be achieved by reusing the existing vocabularies. Reuse will prevent practitioners from creating double concepts. Users should also be able to link them to other semantic artefacts, or integrate them in authoring environments.

Vocabularies should be published as Linked Open Data (i.e. RDF, SKOS, OWL formats) and provide comprehensive coverage of the domain through structured concept definitions and examples. Since SSH is a very diverse domain, concept definitions may vary as well leading to linguistic barriers, such as ambiguity. Therefore, mappings between different vocabularies are needed. However, since it is time-consuming to map entire vocabularies, it was proposed to adopt a flexible semantic mapping framework (e.g. SEMAF EOSC project) targeted at specific interoperability goals.

For the SSH Open Marketplace, vocabularies are essential to describe the entries, improve search and retrieval, and foster discoverability. Furthermore, the external sources ingested into the marketplace will need to be aligned with the "local" vocabularies. Finally, yet importantly, since vocabularies are changing once they are integrated into a platform, a solid governance model is needed to ensure that updates and maintenance are done systematically and automatically. This will ensure quality.

Requirements for vocabulary hosting, management and publication platforms

The milestone report MS8.0 provides a set of recommendations for vocabulary publication platforms for SSHOC. Three out of the eight platforms seem to be the best candidates to host and publish the SSHOC vocabularies, namely ACDH-CH, CESSDA and CLARIN Vocabulary Services. However, none of these platforms seem to fulfil all the requirements as identified in the MS8 report, *Recommendations for vocabulary platforms in SSHOC*.

The panel of the virtual workshop concluded that more discussions are needed to come to a final recommendation for vocabulary platform(s) in SSHOC. Maybe instead of looking for a single platform to host and publish vocabularies, efforts should be made to achieve interoperability between the different platforms at the following levels: data exchange, vocabulary identification schemes, vocabulary maintenance and quality management.

We will use the output of the vocabulary event series to reassess the initial requirements for vocabulary platforms in SSHOC. Last but not least, we will further develop our SSH vocabulary registry, find a suitable platform to store it and investigate how this work could feed into the SSHOC products.

Feedback

Overall, the participants found the event very useful and well-organised. They appreciated the diversity and variety of presentations, the discussion panel and the discussion in the chat.

Here are some testimonials about the impact of the event:

"It will impact my work on the usage of vocabulary to common people. I have learned that I should make the technical words easier to understand to the common people. The more they understand, the more they are likely to cooperate and support."

"It is inspiring to think about Controlled Vocabulary Management."