SSH Vocabulary Initiative - What users want

Laure Barbot[1] , *Daan Broeder*[2], *Matej Durco*[3], *Taina Jääskeläinen*[4], *Iulianna van der Lek*[2], *Monica Monachini*[5], *Irena Vipavc Brvar*[6], *Marieke Willems*[7], *Holly Wright*[8],

[1]dept. name of organization (of Affiliation), name of organization (of Affiliation), City, Country

[1] Digital Research Infrastructure for the Arts and Humanities (DARIAH), Berlin, Germany

[2] Common Language Resources and Technology Infrastructure (CLARIN), Utrecht, The Netherlands

[3] Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH), Vienna, Austria

[4] Tampere University of Applied Sciences , Tampere, Finland

[5] Istituto di Linguistica Computazionale Consiglio Nazionale delle Ricerche (ILC-CNR), Pisa, Italy

[6] University of Ljubljana, Faculty of Social Sciences, Ljubljana, Slovenia

[7] Trust-IT Services, Pisa, Italy

[8] Archeology Data Service, York, United Kingdom

Abstract. SSHOC will build the Social Sciences and Humanities part of the European Open Science Cloud. One of the SSHOC project's core objectives is to foster the transition from the current Social Sciences and Humanities (SSH) landscape to a cloud-based infrastructure that will operate according to the FAIR principles, offering access to research data and related services adapted to the needs of the SSH community. Furthermore, the tools, services, repositories and other resources developed and enhanced in SSHOC will be featured in the SSH Open Marketplace.

The SSH European Research Infrastructure Consortia (ERICs) partnering in SSHOC are exploring and enabling collaboration and deeper integration of each other's infrastructures. One topic that is of relevance for all SSHOC SSH stakeholders is that of managing and using vocabularies. Here we use vocabularies as a general term covering a range of semantic artefacts such as wordlists, taxonomies and thesauri.

The SSH vocabularies are essential for a proper description of resources and phenomena and in SSHOC many tasks are concerned with them. In SSHOC, the SSH Vocabulary Initiative was launched to coordinate related vocabulary activities and investigate, inform and exchange expertise on vocabularies and the platforms that are hosting and managing them.

1. Introduction and Objectives

SSHOC will build the Social Sciences and Humanities part of the European Open Science Cloud. SSHOC has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-04-2018 (GA 823782).

One of the SSHOC project's core objectives is to foster the transition from the current Social Sciences and Humanities landscape to a cloud-based infrastructure that will operate according to the FAIR principles, offering access to research data and related services adapted to the needs of the Social Science and Humanities (SSH) community. Furthermore, the tools, services, repositories and other resources brought in by the project partners or generated during the project will be featured in the SSH Open Marketplace.

The SSH European Research Infrastructure Consortia (ERICs) partnering in SSHOC are exploring and enabling collaboration and deeper integration of each other's infrastructures. One topic that is of relevance for all SSHOC SSH stakeholders is that of managing and using vocabularies. Here we use vocabularies as a general term covering a range of semantic artefacts such as wordlists, taxonomies and thesauri.

The SSH vocabularies are essential for a proper description of resources and phenomena and in SSHOC many tasks are concerned with them. In SSHOC, a specific "Vocabulary Initiative" was launched last year to coordinate related vocabulary activities and investigate, inform and exchange expertise on vocabularies and the platforms that are hosting and managing them. Therefore, the proposed workshop will have the following main objectives:

- To engage the SSH end-user communities present at ICTeSSH in the SSH Vocabulary Initiative, to collect their input and feedback on managing vocabularies, and vocabularies as FAIR semantic artefacts.
- To raise awareness in the SSH research community present at ICTeSSH on finding, understanding and reusing vocabularies via the SSH Open Marketplace.

The workshop took place on 29 June 2021, and was attended by 108 ICTeSSH2021 participants from all over the globe. The slides are available via ZENODO [1] and also the recordings will be made accessible via the SSHOC website [2].

2. The SSH Vocabulary Initiative

The SSH has a considerable large diversity with respect to the studied phenomena, methodologies and data-types used. This is a direct cause for the equally large variety of vocabularies, mostly used for the description of phenomena and data.

The SSHOC project brings together the large European research infrastructure organisations in the SSH, which each have their own experiences and challenges with using and managing vocabularies, and where although there have been smaller collaborations on the topic of vocabularies there was no attempt to come to a common approach.

One of the surprises with respect to the use and work on vocabularies in the SSH as was demonstrated in SSHOC is that there is quite some diversity in available expertise amongst the researchers and infrastructure specialists. It is not always clear within a community who is able to advise and there is certainly no overarching group that has a good overview of the activities in the SSH. Which leads to wasteful actions.

On the other hand, much important work was done, many organizations have good experts and are doing excellent work. And within the SSH, contrary to other disciplines, there is also the work done on the phenomenon of the vocabulary itself beyond the usability or knowledge representation aspects, but looking at vocabularies from a linguistic one.

In SSHOC, investigating a common vocabulary approach was also not planned beforehand. But during the project the realization grew that finding common approaches at the SSH level would be quite useful and some important initiatives were taken. Especially to come to common technical solutions, a number of technical information sessions and accompanying discussions about common vocabulary authoring and management platforms were organised followed by a workshop discussing recommended solutions.

Although those activities were concluded successfully, still more work on SSH vocabularies is thought useful. We have short-listed some important topics where the SSH communities could collaborate delivering important interoperability gains. The current SSHOC Vocabulary interoperability topics are: SSH recommendations for Vocabulary versioning, Vocabulary registration, Vocabulary recommendation.

Some extra effort will be put in further discussions on those topics in the context of the SSH Vocabulary Initiative [3]. Where possible we will look for collaborations with existing initiatives and projects. Feedback from SSH researchers using vocabularies and vocabulary curators and managers is most welcome and we are looking into ways to engage with the wider SSH community on this topic.

3. How can researchers use Vocabulary in SSH tools?

Given the breadth of the Social Sciences and Humanities sector, it is of no surprise that researchers are faced not only with a multitude of theoretical and empirical approaches to research but also with an enormous pool of various tools, systems, and resources that are intended to help researchers in their endeavours. In this section we seek to address how researchers use or can use Vocabulary in SSH tools and provide a few examples in the context of the SSHOC project:

- Using Vocabularies in the SSH community
- Locating suitable vocabularies. How can the SSH Open Marketplace help?
- Interoperability in SSH Vocabulary

3.1. Using Vocabularies in the SSH community

In the context of the SSHOC project, a survey (Petitfils et. al., 2021) [4] was sent out in the first half of 2020 to get a better understanding of the actual use of vocabularies and related work practices in the SSH research communities. The results of the survey were used to inform the SSH open marketplace developments.

Out of 72 complete answers, 52 respondents indicated that their organisations used vocabularies. As the answers were very diverse and a bit biased because most respondents were from France, it was not possible to identify specific trends and practices within the community.

Nevertheless, some insights about the usage of vocabularies were obtained. In general, vocabularies are used to describe disciplines, general concepts, geo entities, persons, and scholarly activities. Due to the insufficient response rate, it was not possible to conclude that

one vocabulary is particularly used within the SSH research community. The vocabularies that stood out were DDI, CESSDA, ELSST, and Pactols [5-8]. This classification of vocabulary types is well known and it will be useful if we set up our own registry.

The vocabularies used by the respondents were collected in a spreadsheet to develop a more elaborate inventory. Each vocabulary was checked to identify the vocabulary type, the format in which it was stored, how it was used by the community, and how often it was updated. The inventory contains about 91 entries and shows that there are many controlled vocabularies of varying scopes, maturity levels and communities using them. BARTOC.org [9] alone currently features almost 3.000 vocabularies from multiple disciplines.

To further investigate the use of vocabularies in the SSH community, a few online information sessions and a workshop were organized where different stakeholders presented their experience with vocabularies and vocabulary management and publication platforms.

Both the survey and the vocabulary online sessions revealed that besides a general lack of knowledge and awareness of vocabularies and their benefits, the following challenges still need to be addressed:

- Some vocabularies might not provide full domain coverage or lack concept definitions and examples.
- Curated vocabularies are not easily reusable when the data model is not the same. This raises barriers for harmonization, search and linking of other vocabularies across domains. CLARIN believes that some of the interoperability challenges could be addressed by reusing existing vocabularies wherever possible, relying on the vocabularies that have already been curated by libraries and linking to authority files wherever possible.
- Availability in translation and authoring environments is limited.

3.2. Locating suitable vocabularies. How can the SSH Open Marketplace help?

One can identify two main uses for vocabularies. Vocabularies are needed to **formalise and conceptualise** a specific dimension or aspect of an application domain. Researchers need them to make the semantics explicit (through verbose descriptions and through relations between concepts). And vocabularies are crucial to build semantic bridges and **interoperability** across projects and dataset boundaries. But their full potential can only be reached if vocabularies are being shared and reused. In that regard, findability of already existing vocabularies is of utmost importance.

To this end a few pointers to vocabulary catalogues:

- The Basic Register of Thesauri, Ontologies & Classifications (BARTOC), as mentioned in section 3.1, contains more than 3.000 vocabularies, but does not include a concept-based index and search.
- Linked Open Vocabularies (LOV) [10] contains around 700 vocabularies, however according to a very broad meaning of the term “vocabularies” (it includes mainly schemas).
- The EU vocabularies website [11][1]provides access to vocabularies managed by the EU institutions and bodies.
- Both SSHOC D3.1 Report on SSHOC (meta)data interoperability problems [12]and D7.6 Resources for Marketplace content description (see section 3.1) have contributed to identify the most used vocabularies within SSH, and a more exhaustive inventory is now being developed.

Despite these different ways and possibilities to locate existing vocabularies, the reuse of vocabularies built by others (usually to serve a slightly different purpose) is not trivial.

Altogether the vocabulary management is a rather complex process as illustrated by the vocabulary's workflow set up at the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) [13]. Between the external authoritative source to be reused, the potential modifications needed in the vocabulary, its publication and persistent archiving, and finally its (re)use in a bespoke application, interconnected infrastructure components and expertise are needed at every step of the workflow.

In practice, the reuse is subject to both technical and conceptual constraints, and is always relative to the need/project at hand. (So the same vocabulary may be perfectly suited for one task, but not for another.) Following aspects need to be considered:

a) it should cover the "right" dimension (i.e. the dimension needed for the specific use-case);
b) it should be comprehensive;
c) stable availability ("cool URIs", stable long-term provision) of the vocabulary, as well as stable reference to concepts should be provided;
d) and it should also be well-established and maintained. The value of a vocabulary is growing with its reuse.
e) It is possible to talk to humans feeling responsible for the vocabulary.

There will probably never be a perfect match between what is needed and what an existing vocabulary is offering. And this is also one of the reasons why it is important to identify the maintainers of a resource, if some adaptation is needed in the existing vocabularies themselves.

The SSH Open Marketplace, built within the SSHOC project can contribute to locate suitable vocabularies and facilitate their reuse. This discovery portal features different kinds of resources - tools & services, training materials, workflows, datasets and publications - useful for SSH researchers looking for digital methods to support their work. One of the central characteristics of the SSH Open Marketplace is contextualisation, i.e. establishing relation between items. These relations will be of great use also for vocabularies: for example, a training material detailing reuse of an existing vocabulary in a tool will be linked to the vocabulary itself, providing useful context for the interested researcher. The concept of this discovery portal is to rely on a crowdsourced curation also supported by research infrastructures communities and experts to ensure that content presented remains up-to-date. The final release of the SSH Open Marketplace is planned for the end of 2021, but a beta version [14], is already accessible.

Based on the SSHOC identification of vocabularies used in SSH - see previous section - dedicated entries for vocabularies will be created in the SSH Open Marketplace. Vocabularies will be referenced as first-class citizens, and the contextualization layer will be added. Let us take the example of the Taxonomy of Digital Research Activities in the Humanities (TADIRAH): an entry is created to reference and describe the vocabulary itself and links & context are providing additional information: the vocabulary repository, the GitHub repository, an interview with the editors of the vocabulary, the publications related to this vocabulary or some links to projects using this taxonomy for example. Thanks to this contextualisation, findability and reusability of SSH vocabularies will be substantially improved.

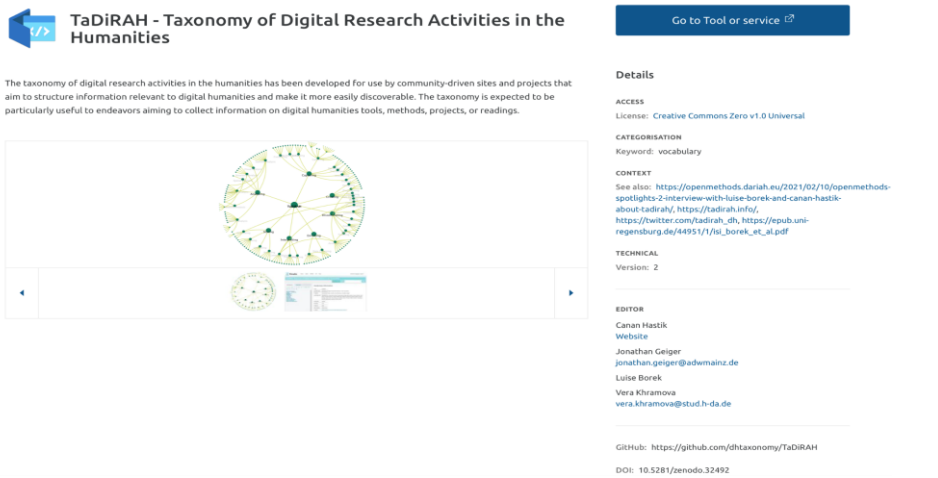

Figure 1: Screenshot of the TaDIRAH entry in the SSH Open Marketplace. **Source:** Development instance, not yet publicly available

3.3. Interoperability in SSH Vocabulary

Interoperability among different terms in a vocabulary is crucial when you work in an international environment, which Europe certainly is.

Multilingual vocabularies are essential for a proper description of resources and phenomena.

This section aims to illustrate the importance of interoperability when working with SSH vocabulary, through the use case of the Vocabulary Matching Tool and examples of machine translation for the creation of multilingual vocabularies.

3.3.1. Vocabulary Matching Tool

ARIADNEplus [15] provides services for archaeologists to enable access to the research infrastructure. In ARIADNE 27 subject vocabularies were identified and a Vocabulary Matching Tool [16] was developed by domain experts from different countries to be able to match local subject terms and concepts to Getty AAT concepts [17]. All the work done in ARIADNE is freely available.

Mapping should be as easy as possible, which leads to continuous work on improvement of interface as well as adding wikidata mapping in order to improve search options. Most of the matching is done manually by subject and language experts and not computers. Experts examine scope and context of the source as well as target concepts before making decisions. While working on making ARIADNE vocabularies FAIR, we make sure that mapping is good enough for others to trust and reuse vocabularies. There is a constant struggle between what should /could be automated and what done manually.

3.3.2. Use of MT for creating multilingual vocabularies

Multilingual vocabularies matter, because they help people find resources and determine their value. Researchers can perform queries in their native language and retrieve data in other languages in the tool, which is especially useful when looking for data or output in non-native languages. Multilingual occupation ontologies [18] could for instance be used and embedded in surveys, so no additional coding and back harmonization is needed. Respondents are able to choose from a detailed list of occupations in male and female form,

which results in more correct entries. The SSHOC project is continuing the work on a multilingual occupational database where all titles are coded according to the International Standard Occupational Classification (ISCO). For the thesaurus European Language Social Science Thesaurus (ELSST) [19] (3.200 concepts), machine translations have been used for two languages to translate the whole thesaurus. This significantly reduced the workload. Machine translation results were reviewed by experts and amended manually if needed. Rough estimate was that about 50% of machine translations were valid as such.

Multilingual vocabularies are especially critical in digital environments, where humans rely on computer processing for reliable and timely results.

4. The SSH Vocabulary Initiative - What Users want

Follow up to the presentation of the topic and introducing work done in different communities and projects, feedback from the audience was sought. The audience was engaged in a series of questions, shared with them a week prior to the workshop, for an agile discussion between the discussion panellists based on the real-time input provided by the workshop participants. The first two ice-breaker questions asked showed that **librarians, researchers & PhDs, and service managers** were the roles most represented among the 23 respondents. Respondents also showed varying levels of familiarity with vocabularies (table 1).

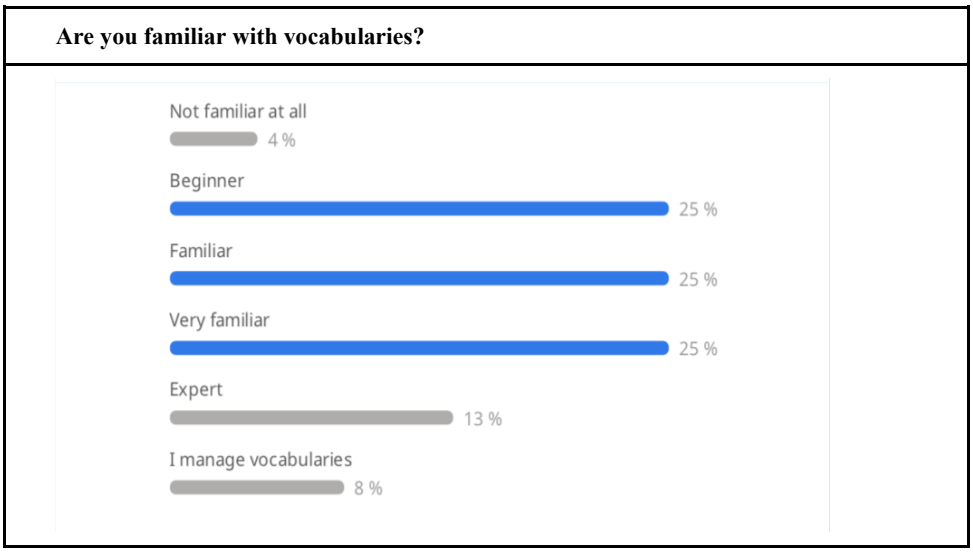

Figure 2: Q1 Are you familiar with vocabularies? Question answered by 24 respondents

4.1. Findability

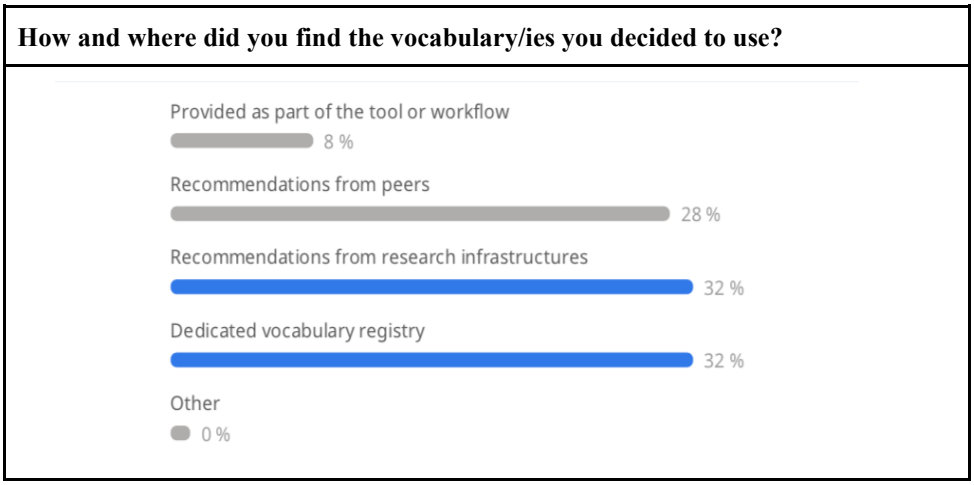

Figure 3: Q2 : How and where did you find the vocabulary/ies you decided to use?Question answered by 25 respondents.

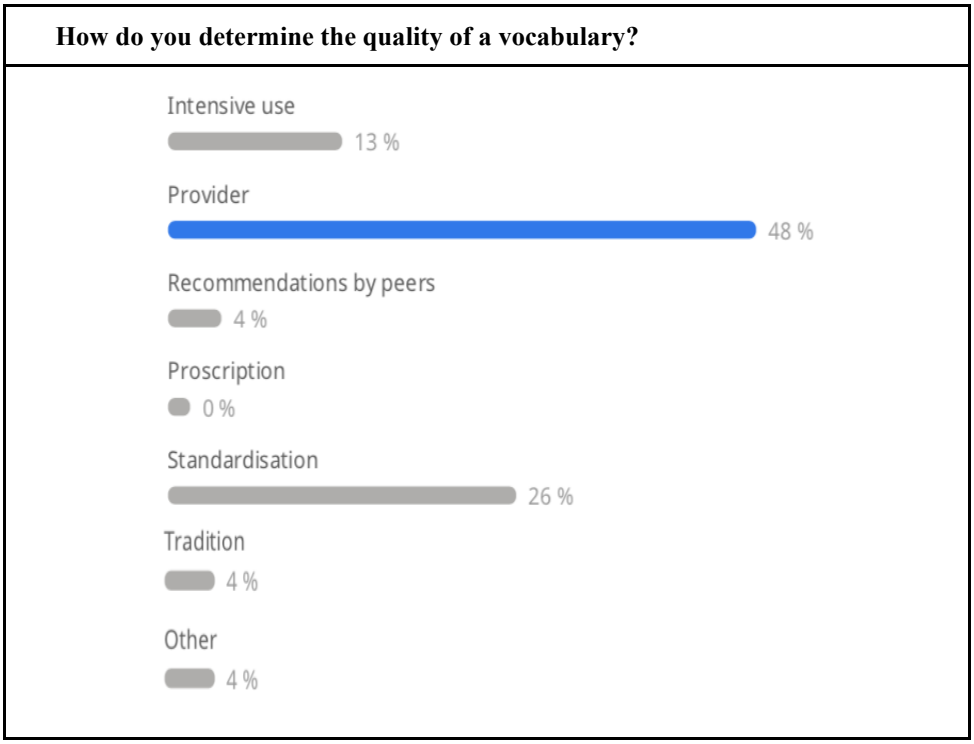

Figure 4: Q3: How do you determine the quality of a vocabulary? Question was answered by 23 respondents.

4.2. Usability

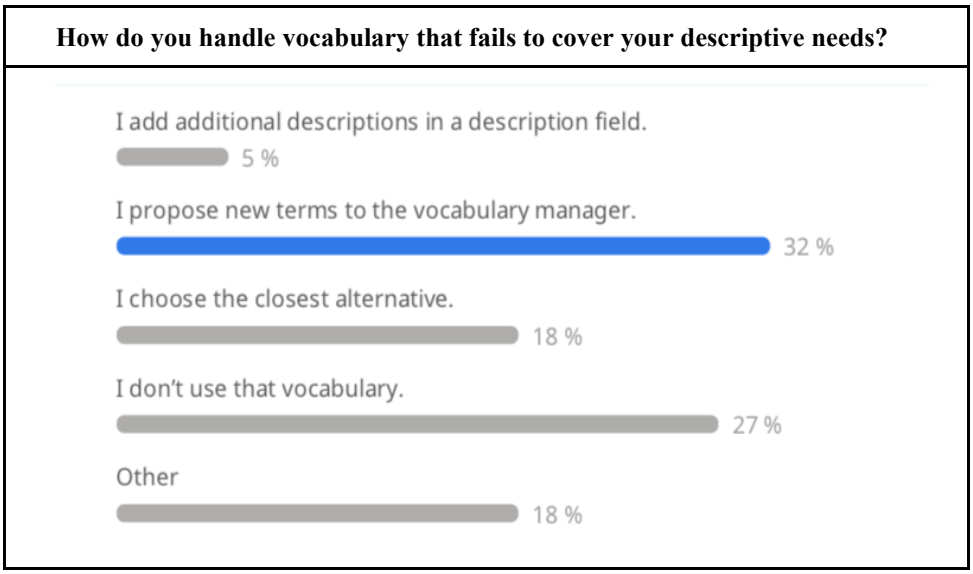

Figure 5: Q4: How do you handle vocabulary that fails to cover your descriptive needs? Question was answered by 23 respondents

Most of the participants mentioned that they would propose a new term to the vocabulary manager in case their vocabulary fails to cover their needs. However, we know that, in practice, accepting a new term could take quite some time.

Another challenge with widely used vocabularies is that whenever there are changes in the vocabulary, the user organisations may need to update their legacy metadata. If they do not, this may have consequences for cross-organisational and multilingual data catalogues, for instance, for filtered searching based on the vocabularies. In addition to good vocabulary management systems, there is also need for good information on the changes made. But perhaps this is part of the workflow?

Governance processes for including new concepts can vary. An example of a governance process that is quite detailed was shared by a participant [20]. A download and adjustment of a vocabulary or even combining several to serve the purpose of the project was mentioned.

4.3. Interoperability

Mapping (conversion) between entries of different vocabularies, how do you manage that?

Figure 6: Q5: Mapping (conversion) between entries of different vocabularies, how do you manage that? Question answered by 13 respondents

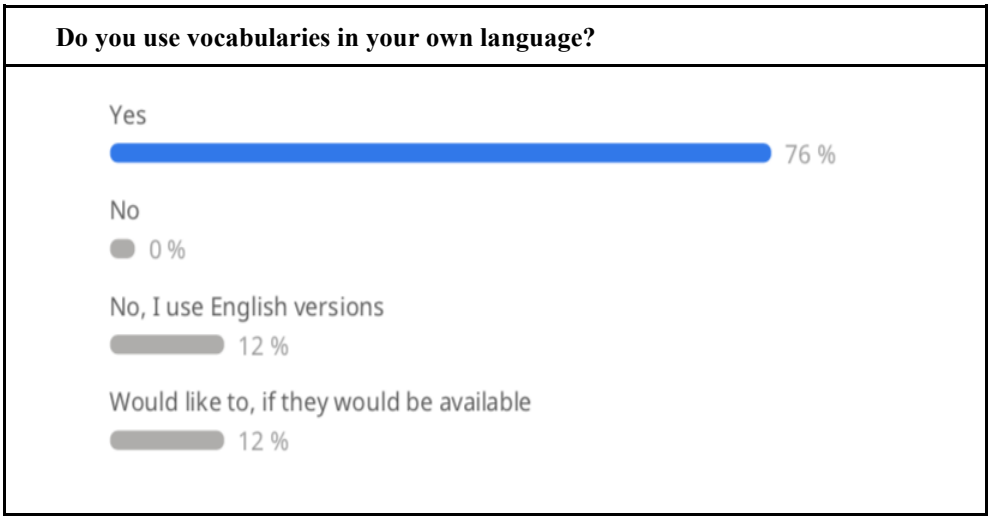

Figure 7: Q6: Do you use vocabularies in your own language? Question answered by 17 respondents

Participants mentioned cross walks and a use of shared spreadsheets, which is especially useful in targeted communities working on the same research project or instrument. However, many did not find a good solution to support their work yet and would be thrilled if the community could offer one.

Several participants would like to use vocabularies in their own language if they would be available, most are already using multilingual vocabularies. Which imply that

multilingualism is an important element in vocabularies and the work being done in this field should continue.

4.4. What users want

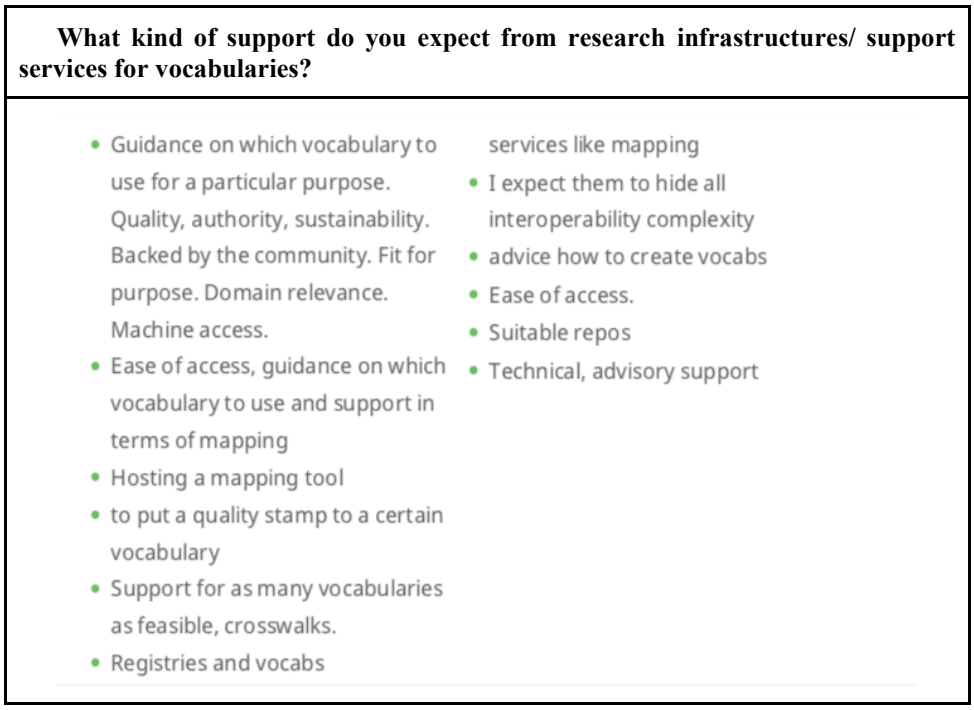

Figure 8: Q7: What kind of support do you expect from research infrastructures/ support services for vocabularies? Question was answered by 11 respondents

Expectations from Research Infrastructures are high, going from sustainability, and support in the way of using the tools and also cross-walks to making sure vocabularies follow FAIR principles and have quality stamp on it. Most are rather challenging, requiring time and resources.

5. Conclusions & next steps

In the workshop we reviewed the uses and needs with respect to vocabularies in the SSH. Here we list the main outcomes from the audience participation and discussion, important for SSHOC work in the SSH Vocabulary Initiative:

1. The appreciation and desire for multilingual vocabularies.
2. Mention of the biomed example for vocabulary registration, e.g. Bioportal and the generic version Ontoportal.
3. The non-technical issues pertaining to vocabulary management/curation of social (involvement) and economic (funding) nature.
4. The importance of the provider when a researcher assesses the quality of vocabularies.

Several research infrastructures and projects are dealing with vocabularies that would cover the needs of their research communities. Some of them were presented at the ICTeSSH2021 pre-conference workshop The SSH Vocabulary Initiative - What Users Want. Expectations of community, also expressed at the workshop, are high and diverse. Where possible the SSH Vocabulary Initiative will look for collaborations with existing initiatives and projects and work towards engaging with wider SSHOC communities on this topic. Feedback from SSH researchers using vocabularies and vocabulary curators and managers is most welcome.

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-04-2018, grant agreement #823782.

References

1. SSH Vocabulary Initiative - What Users Want, ICTeSSH2021 pre-conference workshop slides: https://zenodo.org/record/5045017#.YQQBci2w21t
2. SSH Vocabulary Initiative - What Users Want, ICTeSSH2021 pre-conference workshop promotion and resources [accessed 30 July 2021] https://sshopencloud.eu/events/sshoc-vocabulary-initiative-what-users-want%C2%A0ictessh-2021-sshoc-session
3. Contact mail: sshvocabularyinitiative@sshopencloud.eu
4. Clara Petitfils, Suzanne Dumouchel, Nicolas Larrousse, Laure Barbot, Klaus Illmayer, Matej Ďurčo and Tomasz Parkola. (2021). SSHOC D7.6 Resources for Marketplace content description. *Zenodo.* doi:10.5281/zenodo.4558339 https://zenodo.org/record/4558339#.YNi4nugzZPY
5. The DDI Alliance offers controlled vocabularies, available at [accessed 30 July 2021]:https://ddialliance.org/controlled-vocabularies.
6. The CESSDA vocabulary service contains 28 vocabularies, available at [accessed 30 July 2021]: https://vocabularies.cessda.eu/.
7. European Language Social Science Thesaurus, available at [accessed 30 July 2021]: https://elsst.cessda.eu/.
8. A controlled and shared vocabulary for archeology stored and managed in Opentheso, available at [accessed 30 July 2021]: https://pactols.frantiq.fr/opentheso/.
9. [accessed 30 July 2021] http://bartoc.org/
10. [accessed 30 July 2021] https://lov.linkeddata.es/dataset/lov/
11. [accessed 30 July 2021] https://op.europa.eu/en/web/eu-vocabularies/home
12. [accessed 30 July 2021] https://doi.org/10.5281/zenodo.3569867
13. [accessed 30 July 2021] https://www.oeaw.ac.at/acdh/tools/acdh-vocabularies
14. [accessed 30 July 2021] https://marketplace.sshopencloud.eu/
15. [accessed 30 July 2021] https://ariadne-infrastructure.eu/
16. [accessed 30 July 2021] https://vmt.ariadne.d4science.org/vmt/vmt-help.html
17. [accessed 30 July 2021] http://www.getty.edu/research/tools/vocabularies/aat/
18. [accessed 30 July 2021] https://www.surveycodings.org/
19. [accessed 3 August 2021] https://elsst.cessda.eu
20. [accessed 30 July 2021 https://wiki.earthdata.nasa.gov/display/CMR/Keyword+Review+and+Release+Process