Alignment of multilingual vocabularies in the Social Sciences and Humanities (SSH) Working Group

1. Charter

Controlled vocabularies are a major discovery tool in the digital environment, and this is true also for the Social Sciences and Humanities (SSH). However, while such vocabularies would seem one of the most appropriate tools to ensure a high visibility of the SSH research outputs, the availability, usage, and interoperability of controlled vocabularies remains very heterogeneous. Indeed, the situation regarding controlled vocabularies in the SSH somehow reflects the atomization of the field, characterized by a high scientific, organizational, and technical diversity. A survey conducted during the Social Sciences and Humanities Open Cloud (SSHOC) project thus confirmed that "the SSH vocabularies landscape is still very heterogeneous: the vocabularies used are often very specialized or discipline-oriented"[1]. This means that the SSH community does make use of controlled vocabularies, but "only a few of them are repeatedly used within the research community"[2]. They remain for the most part disconnected from one another, differing not only in scope, but in methodology, accessibility, and level of standardization. From this first set of observations derives the necessity to at least harmonize their discoverability level within the SSH and beyond.

SSH practices related to controlled vocabularies also reflect a specificity of the field that needs to be preserved but poses specific challenges as far as vocabularies are concerned: multilingualism. Research in the SSH can use English as a *lingua franca* for broad dissemination and discussion, but "[r]esearchers from the social sciences and humanities (SSH) who study culture and society often publish in local languages"[3]. In fact, besides the communication dimension, SSH scientific activity is very often rooted in regional rather than global networks and conducted in local languages "because language in those disciplines is not only a communication tool but is entangled into the object of the study"[4]. This aspect obviously also concerns the constitution and use of controlled vocabularies. During a workshop held at the ICTeSSH in 2021 with SSH professionals, a majority of the attendees confirmed that they used controlled vocabularies in their own native language[5]. In that sense, the SSH practices require to have not only high-quality controlled vocabularies, but "high-quality multilingual vocabularies"[6].

This particular dimension of multilingualism has been directly tackled by two recent European projects dedicated to the SSH community: SSHOC and TRIPLE. With distinct but comparable methods, the two projects used machine translation opportunities and human post-validation to generate curated vocabularies in many languages. Among other outputs, the SSHOC project provided a translation from English into 4 additional languages using the CLARIN Concept Registry[7]; the TRIPLE project provided a

[1] Petitfils, Clara, et al. *SSHOC D7.6 Resources for Marketplace Content Description*. Feb. 2021, p.3, https://doi.org/10.5281/zenodo.4558339.
[2] *Ibid*., p.7.
[3] Kulczycki, Emanuel, et al. "Multilingual Publishing in the Social Sciences and Humanities: A Seven-country European Study." *Journal of the Association for Information Science and Technology*, vol. 71, no. 11, 2020, pp. 1371–85, https://doi.org/10.1002/asi.24336.
[4] Leão, Delfim, et al. *OPERAS Multilingualism White Paper*. July 2018, https://doi.org/10.5281/ZENODO.1324025.
[5] Broeder, Daan, et al. *SSH Vocabulary Initiative - What Users Want*, p. 53. https://doi.org/10.5281/zenodo.5045017. Online.
[6] Frontini, Francesca, et al. *D3.9 Report on Ontology and Vocabulary Collection and Publication*. Dec. 2021, p.7, https://doi.org/10.5281/ZENODO.5913484.
[7] *Ibid.*

translation from English into 12 additional languages using the Library of Congress of Subjects Headings[8]. Besides its final result, both activities also outlined the complexity of creating, validating, and maintaining multilingual vocabularies in the SSH context. Indeed, "long-term maintenance and updating represent a further challenge"[9] with vocabularies covering many linguistic areas and cultures and bound to be further expanded.

The RDA WG on 'Alignment of multilingual vocabularies in the Social Sciences and Humanities' intends therefore to address the two main challenges just described of interoperability and multilingualism. It plans to do so by providing collectively approved recommendations for creating, extending, updating, and aligning multilingual vocabularies in the SSH. In that prospect, the SSHOC HORIZON project, and after its end the SSH Open Cluster, have established a roadmap to set-up coordinated actions to support practices sharing, alignment, and federated access to SSH vocabularies. This roadmap for "SSH Vocabulary Commons"[10] converges with the work conducted within the TRIPLE project, as confirmed by a common workshop held by the two entities in 2023.

The objective of this RDA WG, currently co-chaired by OPERAS RI, the coordinator of TRIPLE, is to focus on a part of the SSHOC vocabulary roadmap and to collect from a broader community both insights and feedback about a methodology for the alignment of controlled vocabularies in many languages. As described below, the recommendations will address the quality, development, interoperability, and versioning of multilingual controlled vocabularies in the SSH. Besides SSHOC's and TRIPLE's, additional use cases could be analyzed to establish such recommendations.

Planned deliverables and milestones:

- M6: Methods for creating multilingual vocabularies in the SSH: use case analysis
- M12: Recommendations for creating and extending multilingual and multicultural vocabularies in the SSH
- M18: Recommendations for FAIRifying and versioning of multilingual vocabularies

2. Value Proposition

The creation of a WG within the RDA framework would help to formalise through international collaboration the first phases of the Vocabulary Commons by focussing on essential building bricks. The objective of this group is therefore to work towards the alignment of SSH vocabularies, with a particular focus on topics vocabularies and multilingualism. The main goal is to reach common understanding and practices in creating, updating, and sharing vocabularies. The group plans therefore to publish recommendation(s) for the SSH community to:

- Improve the quality of the controlled vocabularies, by ensuring both the appropriate scientific process and the optimal technical set-up in creating or reusing controlled vocabularies in the SSH;
- Develop the vocabularies' coverage and usage, by sharing experience and knowledge about solutions to address multilingualism and multiculturalism;
- Increase vocabularies' interoperability, by defining the minimum metadata and the representation formats able to ensure the vocabularies FAIR-compliance;

[8] Katsaloulis, Iraklis, and Cezary Rosiński. *How to Create a Social Sciences and Humanities (SSH) Vocabulary: The GoTriple Hackathon Example*. https://doi.org/10.5281/zenodo.5776256. Virtual event.
[9] Frontini, *op. cit.*, p. 33.
[10] Broeder, Daan. *SSH Vocabulary Commons*. Apr. 2023, https://doi.org/10.5281/ZENODO.7871203.

- Facilitate their continuous versioning, by formalising shared rules about concepts and terms updates.

This work will benefit in the first place to the SSH community by providing guidelines that will improve the coverage and the interoperability of multilingual vocabularies. As a first result, it should increase the discoverability of SSH research outputs in the global context, even if produced and described in local languages, by connecting them to many other vocabularies. Additionally, it will benefit the services providing searching tools (libraries, research infrastructures), which will have the possibility to provide accurate descriptors in local languages, while ensuring their connection with descriptors in many other languages. From the point of view of aggregators and, more generally, the global research community, this will increase the diversity and representativity of the results available in search engines.

More specifically, in the prospect of the SSHOC Vocabulary Commons roadmap, such recommendations would concretely help to build in the future a federated access to vocabularies that will be efficiently searchable and interoperable. It would therefore represent an improvement for the SSH researchers, service providers, and end-users, as well as for the whole scientific community and the broader public.

Although different in scope, as only partially focused on the technical framework for semantic interoperability, the group activities will be closely connected to the ones of the Vocabulary Services IG, by increasing the number of semantic artefacts available for more advanced operations.

3. Engagement With Existing Work in the Area

The group gathers at the moment major European Research Infrastructures in the SSH domain (namely, CESSDA, CLARIN, DARIAH, and OPERAS), which have a high expertise in creating and managing controlled vocabularies for a large community of users. It is coordinated by OPERAS, a European Research Infrastructure dedicated to the open scholarly communication in the SSH, as part of its contribution to one of the pilots of the RDA TIGER project.

As mentioned in the Charter above, the WG is a follow-up of both SSHOC and TRIPLE projects and is based on theirs and other researchers' reports. A Zotero group will collect the relevant references on the topic, often mentioned, but not yet fully addressed. It is furthermore expected that opening the discussion into RDA will bring some useful insights and suggestions from outside Europe and the Western area, as the English as the scientific *lingua franca* can be even more challenging beyond this area.

4. UN Sustainable Development Goals (SDGs)[11]

The aims of the WG align with two of the UN SDGS:

[11] The SDGs currently use a controlled vocabulary maintained in the SDMX Global Registry: https://registry.sdmx.org/items/conceptscheme.html. SDMX is an international standard used by the official statistics community for exchanging data and it allows terms to be defined in multiple languages. It could be therefore a very useful resource for the WG.

- **Goal 4 “Quality education”:** The WG will increase the quality of research outputs dissemination, bringing more and more diverse knowledge to the scientific community, but also to the broader community.

- **Goal 11: “Reduced inequalities”:** By focusing on the SSH fields, which can have great social impact, and on both multilingual and multiculturalism, the WG will improve the level of diversity of the available research outputs, thus slightly amending the unequal representation of countries, regions, and forms of knowledge.

5. Work Plan

1. Final recommendation of the WG

The final recommendation will address the methodology to create and update multilingual vocabularies in the SSH and the best practices to FAIRify the vocabularies and their different versions. It will present the methodology’s main steps and technologies that should be used to create and update multilingual vocabularies. It will also describe how to make vocabularies FAIR and citable in all their versions.

2. Milestones and intermediate outputs

a. M6: “Methods for creating multilingual vocabularies in the SSH: use case analysis”: This milestone will be a report on the various methodologies used to create and update multilingual vocabularies, starting from, but not limited to, the examples of SSHOC and TRIPLE.

b. M12: “Recommendations for creating and extending multilingual and multicultural vocabularies in the SSH”: This deliverable will give the basis for an intermediate deliverable that will provide recommendations on the methodologies.

c. M18: “Recommendations for FAIRifying and versioning of multilingual vocabularies”: During the last phase, the WG will focus on a deliverable dedicated to more specific aspects of vocabularies FAIRification and versioning.

d. M18: “Final recommendations about SSH multilingual vocabularies alignment”: The final recommendations of the WG will consist of the merging of the two previous deliverables.

3. WG’s activities organization

The meetings and collective work will be exclusively online.

4. WG’s coordination and monitoring

The WG entails in its objectives the inclusion of less represented disciplines, languages and cultures, and as such relies on creating a broad consensus.

Furthermore, the leading members of the WG are used to collective and multicultural work through both their institutions and contributions in international projects, and are therefore experienced in terms of reaching collectively approved outputs.

The experience of the members in leading and contributing to research projects will ensure the definition of reachable objectives and their achievement in due time.

5. Community engagement

The WG already represents an important part of the SSH community in the European context, through the SSH Open Cluster and the contribution of SSH ERICs. The members are also involved in other WGs or task forces within RDA, EOSC and their own professional networks. All these communication channels will be used to ensure a major impact of the recommendations.

6. Adoption Plan

The first adoption step should take place within the frame of SSHOC and TRIPLE, by aligning methodologies, FAIRification and versioning practices amongst the members of the WG and their respective organizations.

The WG will also give the opportunity to identify additional use cases in the global SSH community where the recommendations could be applied. It is in particular expected that the co-chairs provide such use cases from outside Europe.

The promotion of such adoptions and their benefits for the whole SSH community through the members' network will allow for broader adoption of the recommendations.

7. Initial Membership

Co-chairs:
Arnaud Gingold (OPERAS-OpenEdition), Gema Bueno-de-la-Fuente (University of Zaragoza), Terhi Nurmikko-Fuller (Australian National University)

Members:

Name	Institution
Ciprian Gerstenberger	Norwegian Centre for e-Health Research
Pascal Flohr	DANS-KNAW
Jetze Touber	DANS-KNAW
George Alter	University of Michigan
Daniela Rosu	University of Toronto

Yukio Maeda	University of Tokyo
Douglas Tudhope	University of South Wales
Martin Benjamin	Kamusi Project International
Janez Štebe	University of Ljubljana
Maja Dolinar	University of Ljubljana, Slovenian Social Science Data Archives
Cesare Concordia	ISTI-Italian National Research Council (CNR)