Research and Innovation Action

Social Sciences & Humanities Open Cloud

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

Deliverable D7.6

Resources for Marketplace content description

Dissemination Level	PU
Due Date of Deliverable	31/12/2020 (M24)
Actual Submission Date	15/02/2021
Work Package	WP7 Creating the SSH Open Marketplace
Task	Task 7.4 - Governance: Population, Curation & Sustainability of the SSH Open Marketplace
Type	Report
Approval Status	Waiting EC approval
Version	V1.0
Number of Pages	p.1 – p.45

Abstract:

This deliverable offers an analysis based on 330 answers to the survey conducted within Task 7.4 of SSHOC project. The survey aimed to understand common practices regarding SSH vocabularies among the research community. The objective was threefold: (1) to provide an analysis on the use of SSH vocabularies within the dedicated research community, (2) to support the implementation of the SSH Open Marketplace through the realisation of the vocabularies' mapping process and the description of resources, and (3) to constitute a complementary study in the light of other involved partners and Horizon 2020 projects.

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

History

Version	Date	Reason	Revised by
0.0	06/2020	Table of content suggestion	Clara Petitfils (CNRS/Huma-Num)
0.1	09/2020	Table of content reviewed by all partners involved	Clara Petitfils, Suzanne Dumouchel (CNRS/Huma-Num), Laure Barbot (DARIAH), Klaus Illmayer (OEAW), Tomasz Parkola (PSNC)
0.2	18/11/2020	First draft	Clara Petitfils (CNRS/Huma-Num)
0.3	04/12/2020	Final draft	Clara Petitfils, Suzanne Dumouchel, Nicolas Larrousse (CNRS/Huma-Num), Matej Durco (OEAW), Laure Barbot (DARIAH), Stefan Buddenbohm (UGOE)
0.4	09/12/2020	Deliverable sent to reviewers	Mélanie Bunel (CNRS/Huma-Num), Iulianna van der Leck (CLARIN)
0.5	16/12/2020	Deadline for reviewers' comments	Mélanie Bunel (CNRS/Huma-Num), Iulianna van der Lek-Ciudin (CLARIN)
0.6	16/12/2020 - 18/12/2020	Inclusion of reviewers' remarks and comments	Clara Petitfils (CNRS/Huma-Num), Laure Barbot (DARIAH)
1.0	02/02/2021	Final version edited and ready for submission	Clara Petitfils (CNRS/Huma-Num), Laure Barbot (DARIAH)

Authors List

Organisation	Name	Contact Information
CNRS (Huma-Num)	Clara Petitfils Suzanne Dumouchel Nicolas Larrousse	clara.petitfils@huma-num.fr suzanne.dumouchel@huma-num.fr nicolas.larrousse@huma-num.fr
DARIAH	Laure Barbot	laure.barbot@dariah.eu
DARIAH/OEAW	Klaus Illmayer Matej Ďurčo	Klaus.Illmayer@oeaw.ac.at Matej.Durco@oeaw.ac.at
DARIAH/UGOE	Stefan Buddenbohm	buddenbohm@sub.uni-goettingen.de
DARIAH/PSNC	Tomasz Parkola	tparkola@man.poznan.pl

Executive Summary

This document presents the results of a survey on existing vocabularies in Social Sciences and Humanities (SSH) conducted by Task (T) 7.4 of the Social Sciences and Humanities Open Cloud (SSHOC) project funded by the European Commission under Grant #823782. T7.4 is part of WP7, in charge of creating the SSH Open Marketplace: a discovery portal crafted for the SSH research community, offering contextualised solutions to diverse research enquiries (e.g., tools or services) at every step of a researcher's data life cycle.

The main objective of this deliverable is to contribute to a better understanding of vocabularies used within SSH in order to inform SSHOC services' development: the SSH Open Marketplace but also other initiatives such as the work conducted by CLARIN and SSHOC T3.1 to improve the service offered by vocabulary management and publication platforms for SSHOC, and the SSHOC Reference Ontology (T4.8).

The survey presented here, distributed among SSH research communities in Europe, gathered 330 answers, with 42 of them commonly using vocabularies, embedded in their research practices.

As the conclusion shows, the SSH vocabularies landscape is still very heterogeneous: the vocabularies used are often very specialised or discipline-oriented, which limits their potential for reuse over the SSH community at large and does not make them very suitable for the generic description needed in a discovery portal such as the SSH Open Marketplace.

This deliverable and the survey also illustrate the diversity of understanding that lies behind the term "vocabulary". Collecting and showcasing the semantic artefacts in the SSH Open Marketplace - be they controlled vocabularies, dictionaries and thesauri to metadata schemas, and ontologies - would be a contribution towards addressing the challenges that are raised by this vocabulary topic in SSH. Alongside the CLARIN and T3.1 initiative to identify a suitable vocabulary platform for editing and publishing SSH vocabularies, and the different tasks working to further develop semantic artefacts for their communities, this could be seen as a significant SSHOC contribution.

Abbreviations and Acronyms

CIDOC-CRM	CIDOC Conceptual Reference Model
CESSDA	Consortium of European Social Science Data Archives
CLARIN	Common Language Resources and Technology Infrastructure
DC	Dublin Core
DDI	Data Documentation Initiative
ELSST	European Language Social Science Thesaurus
EOSC	European Open Science Cloud
ISO	International Organization for Standardization
MORESS	Mapping of Research in the European Social Sciences and Humanities
MS	Milestones
ORCID	Open Researcher and Contributor ID
OSPP	Open Science Policy Platform
RDF	Resource Description Framework
SKOS	Simple Knowledge Organization System
SSH	Social Sciences and Humanities
SSHOC	Social Sciences and Humanities Open Cloud
SSHOCro	SSHOC Reference Ontology
T	Task
TaDiRAH	Taxonomy of Digital Research Activities in the Humanities
TRIPLE	Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration
WP	Work Package

Table of Contents

1. Introduction 6

2. Methodology 9

The questionnaire 9

Data Analysis, Communication and Distribution 11

3. Analysis and Results 13

Respondents' Profile 13
Respondents' SSH disciplines 13
Respondents' organisation type 15
Respondent's country of residence 16

On the use of vocabularies 16
On the use of SSH vocabularies and their identification 16
On the areas, discipline specificity and languages of vocabularies 17
On vocabularies' alignments 18
On vocabularies' obsolescence, self-creation and online availability 19

Concluding remarks 20

4. Relations to the SSHOC project 22

WP7's work on the SSH Open Marketplace mapping process 22

SSHOC's WP3 activities on vocabularies' use consolidation 23

Links with other SSHOC initiatives: the SSHOC Reference Ontology (SSHOCro) 24

5. Conclusion 26

6. References 27

Annex 1 - The questionnaire 28

Annex 2 - Respondents' research organisation / infrastructure 31

Annex 3 - On the use of SSH vocabularies 33

Annex 4 - Vocabularies' alignment 40

Annex 5 - Expectations for vocabularies' alignment 42

Annex 6 - Online vocabularies' URLs 43

1.Introduction

This deliverable is part of the SSHOC project's Task 7.4 "Governance: Population, Curation & Sustainability of the SSH Open Marketplace" led by CNRS/Huma-Num. It is a component of WP7 "Creating the SSH Open Marketplace". Its objectives are threefold: (1) provide an analysis on the use of SSH vocabularies within the dedicated research community, (2) support WP7's goal to create an SSH Open Marketplace and (3) constitute a complementary study in the light of other SSHOC involved partners and Horizon 2020 projects.

Better understanding of researchers' practices

Vocabularies[1] - especially controlled vocabularies - are playing a crucial role in offering a structured and coherent description of the content, and in increasing the findability of resources. The more known and reused the vocabularies are, the more useful and meaningful they become, thus allowing better semantic interoperability between systems.

Alongside other SSHOC initiatives, the team of the Task 7.4 in SSHOC conducted a survey to better understand the common use and practices throughout the European research community - especially through the second set of questions. Or, in other words: **what are SSH researchers' most common practices regarding vocabularies?** For instance, the team attempted to understand which vocabularies were the most commonly used, which languages were favoured or what was the alignment envisioned and expected from the SSH research community.

This understanding and mapping of the SSH research communities' practices is essential to identify researchers' most common practices and thus potential gaps, not only for future studies but also for the creation of the SSH Open Marketplace's.

Supporting the development of the SSH Open Marketplace

One of the goals of this survey was to inform the development of the discovery portal "SSH Open Marketplace", the primary output of WP7. The Marketplace shall provide information about tools and other resources as well as workflows relevant for the SSH community in the form of a searchable and browsable catalogue. To this purpose, information about relevant resources is being ingested from existing third-party sources and is mapped to a marketplace-specific data model that allows describing these resources in a way, which would allow for easy discovery of the relevant resources by interested users. As with any catalogue, controlled

[1] Vocabularies are here understood according to two definitions: "Any prescribed set of terms used to limit the acceptable set of metadata values can be referred to as a vocabulary" (Hider 2012, p. 151) and "A vocabulary is a computer-readable file that captures terms, their URIs, and descriptions" (Go Fair principles) #.

vocabularies provide a consistent way to describe and structure data, improving findability of the resources and information retrieval.

Therefore, several descriptive properties have been identified, which should use concepts from a controlled vocabulary as values. The more established and well-known these vocabularies are, the more useful they are because the users are better acquainted with them and thus better understand their structure and the meaning of individual concepts. Moreover, reuse of vocabularies is crucial for achieving semantic interoperability between systems and datasets. One could even argue that controlled vocabularies are only used to their full potential if they are being reused. Therefore, the T7.4 team aimed at surveying the landscape for existing established vocabularies[2], which could be employed in describing the resources in the Marketplace. In general, T7.4 found that the landscape is quite heterogeneous: many vocabularies are being used, only a few of them are repeatedly used within the research community. Also, many of the vocabularies are quite specialized and thus not suitable for the rather general description of items in the Marketplace.

However, even if individual vocabularies cannot be used in the Marketplace data model, they still can and should be added to the Marketplace as individual semantic artefacts, as items of the Marketplace. Towards this end, the survey offers a valuable comprehensive inventory to source from. In this respect it is worth mentioning that the term "vocabulary" itself seems to be interpreted very broadly and differently by individual respondents, subsuming under it, all kinds of semantic artefacts ranging from actual controlled vocabularies, dictionaries and thesauri to metadata schemas, and ontologies. Collecting and systematically structuring all these different resources in the Marketplace would be a contribution to clearer distinction and hopefully, a better understanding of this area within the community.

A complementary initiative for other partners and projects

The potential complementarity of this analysis is important to consider, first in the light of the SSHOC project. Indeed, throughout the project, it was noted that several similar initiatives were ongoing. For instance, WP3 "Lifting Technologies and Services into the SSH Cloud" led by CLARIN ERIC included the following vocabulary-related tasks:

- Task 3.1 "Multilingual Terminology" led by CNR that will provide the deliverable D3.9 *Report on Ontology and Vocabulary collection and Publication* next year, in M36.
- Task 3.2 "Selected SSH Ontologies and Vocabularies" led by CLARIN ERIC that aims to "*foster the use of selected global ontologies in the SSH, regarding occupational titles, educational categories, sector of the industry, geographical regions, food items, and religions*" through deliverables D3.4 *Multilingual ontologies for Occupation, Industry, Regions and cities, Food items, and Religion, with use cases* due for M26 and D3.10 *A 20th-century version of the occupation multilingual ontology* for M38.

[2] A first list of 15 potential vocabularies that could be used to describe Marketplace content was provided in the previous deliverable D7.1, https://doi.org/10.5281/zenodo.3547648, see table 23 "Vocabularies applicability in the Marketplace": CESSDA Topic Classification, CLARIN Concept Registry, ISO 639 - 1 language list or TADIRaH were for example mentioned there.

- Task 3.5 "Data and Metadata Interoperability Hub" with D3.1 *Report on SSHOC (meta)data interoperability problems*[3] released in 2019.

In this context, CLARIN ERIC took the initiative to try harmonising all the SSHOC vocabularies-related work. This harmonisation exercise translates, among others, in the organization of a dedicated Workshop "SSHOC Consideration for the Vocabulary Platforms" on November 6th, 2020. This deliverable is therefore part of this ongoing exercise.

A partnership has been set up with another Horizon 2020 project: the TRIPLE project[4], coordinated by CNRS/Huma-Num. Since CNRS/Huma-Num coordinates both the TRIPLE project and SSHOC's Task7.4, it was important and natural to think about potential synergies. The outcome of this deliverable will therefore be useful for TRIPLE's WP2 "Data acquisition and categorisation," which is in charge of defining the technical specifications of data acquisition for TRIPLE's search engine. The result of the survey could also help the TRIPLE project to identify vocabularies for the description of the content. Overall, it is expected that this deliverable will lay the basis for future work on vocabularies and their alignment. Moreover, organising and structuring the use of vocabularies in the SSH communities could ultimately help improve the interoperability across platforms.

[3] Broeder, Daan, Trippel, Thorsten, Degl'Innocenti, Emiliano, Giacomi, Roberta, Sanesi, Maurizio, Kleemola, Mari, … Ďurčo, Matej. (2019). SSHOC D3.1 Report on SSHOC (meta)data interoperability problems (Version v1.0). Zenodo,.https://doi.org/10.5281/zenodo.3569867, accessed 18 December 2020.

[4] European Commission, Transforming Research through Innovative Practices for Linked interdisciplinary Explorationhttps://cordis.europa.eu/project/id/863420, accessed December 16th, 2020.

2. Methodology

This section presents the methodology adopted to conduct this survey from the questionnaire drafting to its dissemination and analysis of the results.

The questionnaire

To answer the survey's threefold objective, thorough preparatory work was necessary. A dedicated Task Force was established within T7.4 to draft the first version of the questionnaire and collect feedback. During the drafting phase, the Task Force encouraged the re-use of previous materials: either from the SSHOC project itself (deliverables and milestones) or from previous and ongoing European projects and initiatives.

First, it was important to offer a clear definition of what the term "vocabulary" encompasses. To prevent confusions between taxonomies, (controlled) vocabularies and ontologies, it was necessary to provide the respondents with a clear definition. For this, the Task Force relied on WP3's work on vocabularies, namely D3.1 *Report on SSHOC (meta)data interoperability problems*[5] and MS8 Report *Choice of vocabulary publication platform for SSHOC*. Definitions were reused from those written outputs and completed by a definition coming from the GoFair initiative.

Two definitions of vocabulary were therefore provided to respondents:

1. "Any prescribed set of terms used to limit the acceptable set of metadata values can be referred to as a vocabulary" (Hider 2012, p. 151)[6]
2. "A vocabulary is a computer-readable file that captures terms, their URIs, and descriptions" (GoFair principles)[7].

The definition being settled, it was important to determine the audience targeted for this study. The choice was based on WP7 previous work, and especially D7.1 *System specification - SSH Open Marketplace*'s suggested groups of the targeted audience: researchers and research supporting staff. To deepen the respondents' background classification, the Task Force exploited some categorisations developed under the MORESS

[5] Broeder, Daan, Trippel, Thorsten, Degl'Innocenti, Emiliano, Giacomi, Roberta, Sanesi, Maurizio, Kleemola, Mari, … Ďurčo, Matej. (2019). SSHOC D3.1 Report on SSHOC (meta)data interoperability problems (Version v1.0). Zenodo,.https://zenodo.org/record/3569868#.X7zVj6pKhQI, accessed 18 December 2020.

[6] cf. SSHOC D3.1; original source: Hider, Philip (2012). Information Resource Description. London: Facet Publishing. https://books.google.de/books?id=jGpVPgAACAAJ&lpg=PP1&pg=PA151#v=onepage&q&f=false

[7] GO FAIR, FAIRification Process https://www.go-fair.org/fair-principles/fairification-process/ accessed 17th December 2020

project[8] for research expertise, especially the 27 SSH disciplines[9] classification. Regarding the research-related profession of the respondents, T7.4 opted for the agreed SSHOC stakeholders categories[10], mentioned both in D2.1 *SSHOC Overall Communication and Outreach Plan*[11] and D6.1 *Community Engagement Strategy*. As explained in D6.2, the project partners identified the relevant stakeholder groups for the SSHOC project in the project proposal stage so that engagement could begin as soon as possible. The selection of the stakeholder categories was informed by the main EOSC stakeholders as identified by the EOSC pilot project, the recommendations of the Open Science Policy Platform (OSPP) and insights from SSHOC partners. Adjustments have been made, considering the developments within EOSC.

Finally, regarding the language of the survey, it was commonly decided that English would be chosen to avoid any confusion of translation and then of interpretation - especially regarding the different similar terms "vocabulary", "ontology" that can convey another definition in a different language.

The questionnaire was divided into three sets of fourteen questions - see Annex 1 for the complete questionnaire. Both open-ended questions and closed questions were chosen.

The first set "Profile" was composed of three questions about the respondents' research background to help us determine the following:

1) The research discipline (following the MORESS classification);
2) The research activity type through the respondents' research organisation type (following the SSHOC stakeholders categories);
3) The country of residence.

The second question of this first set was organised into two parts: (1) respondents were asked to indicate the name of their research organisation and (2) to describe it according to this classification:

- Researchers
- Research Libraries and Archives
- Research & e-Infrastructures & EOSC thematic clusters
- Universities and Research Performing Organisations
- Research Funding Organisations
- Policy-Making Organisations
- Private Sectors and Industry Players

[8] CORDIS EU research results. (2005) Mapping of research in European social sciences and humanities. https://cordis.europa.eu/project/id/HPSE-CT-2002-60060, accessed December 18th 2020.

[9] The 27 disciplines described by the MORESS project are used by the French Open Archives HAL with the available list here: https://halshs.archives-ouvertes.fr/browse/domain

[10] SSHOC. Who benefits. https://www.sshopencloud.eu/who-benefits, accessed December 18th 2020

[11] SSHOC. D2.1 SSHOC overall Communication and Outreach Plan https://www.sshopencloud.eu/d21-sshoc-overall-communication-and-outreach-plan, accessed December 18th 2020

- Civil Society and Citizen Scientists
- Others.

The second set contained ten questions about the use of vocabulary in the SSH research community. The questions were organised into different themes: languages, online availability, alignment or self-creation.

For the fifth question of this second set, ("What is/are the vocabulary(ies) used for? Which area/topic is it covering?") respondents could choose:

- Named entities: persons
- Named entities: geographical entities
- Named entities: institutions
- Concepts: Scholarly activities
- Concepts: Disciplines
- Concepts: General concepts
- Others: please specify.

Finally, the third set of questions related to the community engagement activity. Respondents were invited to participate in the SSHOC activities either through a follow-up group on vocabularies or by subscribing to the SSHOC Newsletter Membership. About 35 respondents showed interest in the follow-up group (be it Task Force or working group). The goal of this group which will be later determined and shaped according to the needs. It nonetheless remains a useful source of information on vocabularies' users in SSH. T7.4.

Data Analysis, Communication and Distribution

To support both the dissemination of the survey and its analysis, the questionnaire was created with LimeSurvey[12] (provided by DARIAH/OEAW). With the help of LimeSurvey's statistical functions, T7.4 was able to retrieve the required results. Those results were extracted based on the survey's objectives: (1) deliver an overview of the SSH research community practices in terms of vocabularies, (2) support the SSH Open Marketplace's implementation and (3) provide an additional study and support to other ongoing and similar activities within the SSHOC and TRIPLE projects.

Therefore, to comply with those objectives, T7.4 Task Force attempted to extract the exact result for each question. The different results' scope offered through LimeSurvey were taken into account: between all answers considered, incomplete answers only or complete answers only. As it is further explained below, each result was carefully extracted and the choice to select one answers' scope over another one also justified.

Regarding the dissemination and communication of this survey, T7.4 pursued a close collaboration with SSHOC's WP2 - "Communication, Dissemination and Impact", led by TRUST-IT. The survey was channelled

[12] See: https://www.limesurvey.org/en/

through WP7 and WP2 to various communication networks and communities from February to May 2020. The TRIPLE project's team also contributed to the communication efforts and distribution of the survey.

3. Analysis and Results

From February to May 2020, **330 answers** were collected for this survey. Out of those 330 answers, **258** were **incomplete** and **72** were **complete**.

First, a survey dropout analysis was performed to understand why and at which stage respondents quitted the survey without completing it. Almost 31% of all respondents ("complete" and "incomplete" answers) quit the survey after the first set of questions, relating to the respondent profile and more precisely to its first question ("What is your SSH discipline?").

About 89% of the respondents who did not complete the survey stopped at question 4 ("Do/Does you/your organisation use vocabularies?"). Out of the total number of respondents (330), only 52 (about 15.7%) confirmed that their organisation used vocabularies. If we take into consideration only the ''complete answers" (72), only 42 (about 58%) replied positively to this question.

The survey dropout analysis suggests that some respondents stopped completing the survey because of the following reasons:

- Respondents were not sure about their SSH discipline.
- Respondents did not understand what it was meant by ''vocabularies.''
- Respondents stopped completing the survey because their organisation did not use vocabularies.

The following section presents the results of this survey. The results have been rounded up to the first decimal for the sake of clarity.

Respondents' Profile

This section presents the results of the first set of questions. The objective of those results' extraction was to identify the main trends of respondents' profiles but also the potential gaps for future studies. The reader should be aware at this stage that no specific trends nor practices could be highlighted.

For each question - and when relevant - only the top four answers were taken into consideration because no relevant results/categories stood out after the fourth result/category.

Respondents' SSH disciplines

The 330 respondents who attempted to answer the survey represented three disciplines: **Linguistics, Archaeology & Prehistory and Sociology**. Here is an overview of the breakdown:

- 15% for Linguistics
- 11.4% for Archaeology & Prehistory
- 10,8% for Sociology

- 7.8% for History.

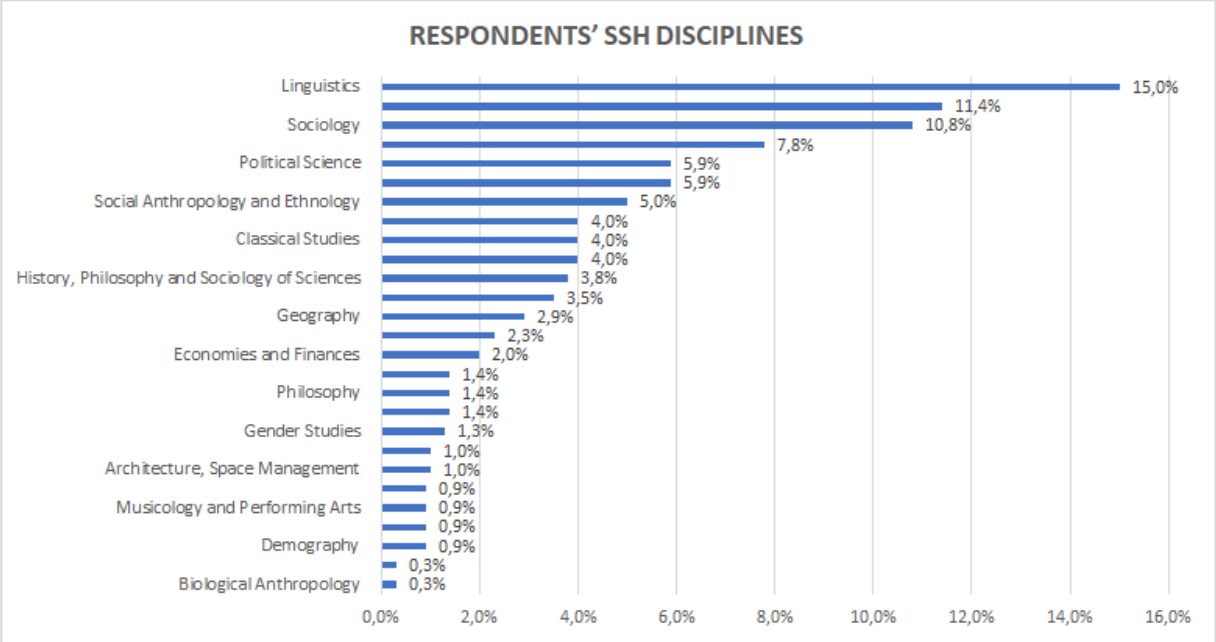

Figure 1. Respondents' SSH disciplines all answers considered

When considering only the "complete" answers, the top three discipline are the same, while the fourth category (History) is replaced by Communication Sciences:

- 26,4% for Linguistics
- 20.8% for Archaeology & Prehistory
- 16.7% for Sociology
- 13.9% for Communication Sciences.

All answers considered, the respondents who indicated that their organisation used vocabularies represented the following disciplines:

- 32.1% for Archaeology & Prehistory
- 28,3% for Linguistics
- 17% for History, Philosophy and Sociology of Science
- 15.1% for Art & Art History.

For this last selection one can note that, although Archaeology & Prehistory and Linguistics are still represented, two new disciplines appear as well: **History, Philosophy and Sociology of Science and Art & Art History**.

Respondents' organisation type

For the first part of this question, CNRS is the most mentioned research organisation among respondents with 18 repetitions (see the full list in Annex 2).

Regarding the organisation type, the following results stood out, all scopes considered (all answers, incomplete, complete): **Universities and Research Performing Organisations, Researchers** and **Research Libraries & Archives**. Here is a visualisation of the breakdown overview:

- 32.9% for Universities & Research Performing Organisations
- 20% for Researchers
- 3.9% for Research Libraries & Archives
- 1.5% for Research & e-Infrastructure and EOSC thematic clusters.

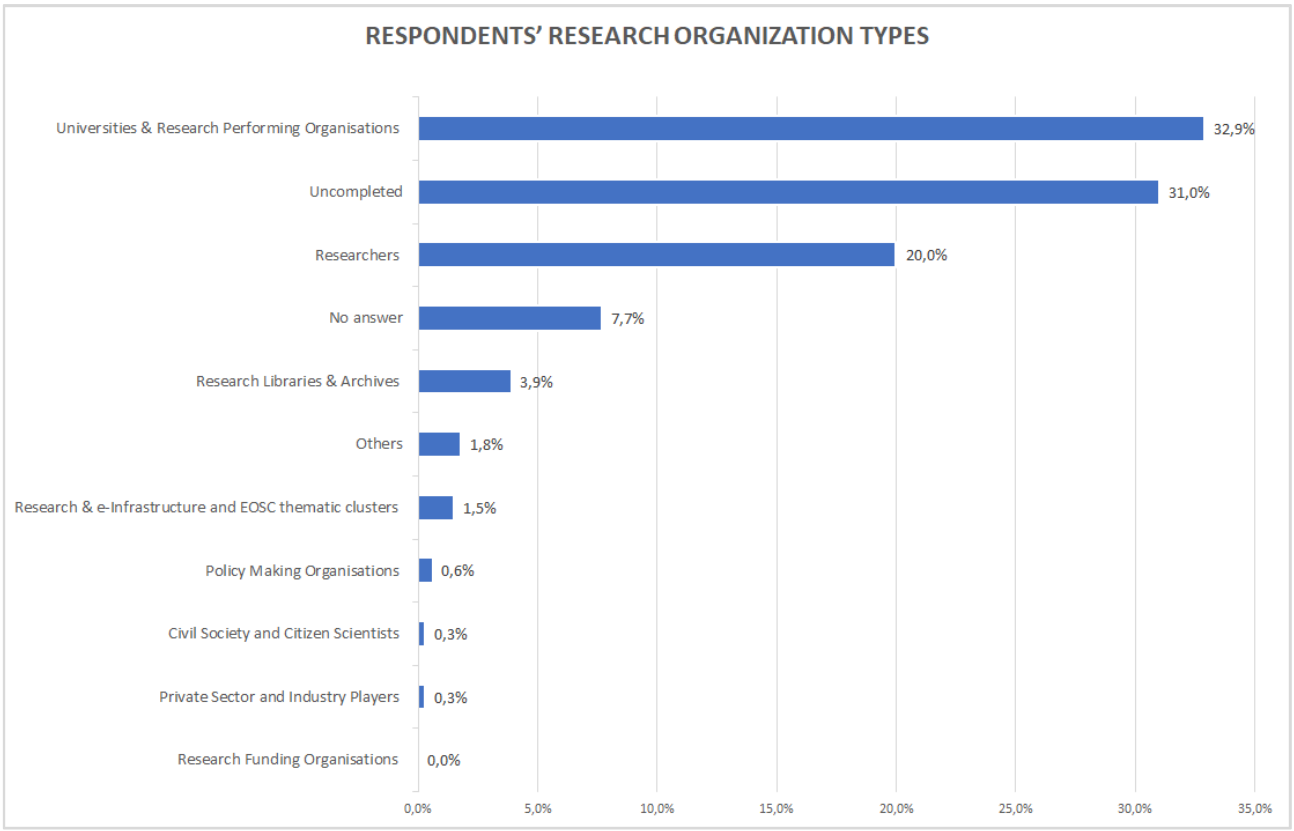

Figure 2. Respondent's research organisation types, all answers considered

When considering either the "complete" answers or respondents' who answered positively to the fourth question, the "**Policy-Making Organisations**" category appears in the top 4. For instance, for complete answers, here is the ranking:

- 62.5% Universities & Research Performing Organisations
- 20.8% Researchers
- 8.3% Research Libraries & Archives
- 2.78% for Policy-Making Organisations.

Respondent's country of residence

Regarding the fourth question, whatever the scope of the analysis was (all answers considered), it appears that most of the respondents were French residents. For instance, regarding complete answers only, the following results show:

- 55,6% from France
- 5.6% from The United Kingdom, Spain and The Netherlands
- 4.2% from Germany.

On the use of vocabularies

The second part of the questionnaire analyses the respondents' use of SSH vocabularies. Only the complete answers are taken into consideration.

On the use of SSH vocabularies and their identification

Respondents were asked whether they used any vocabularies, to specify which ones (with ten available blank entries) and also, in case of a negative answer, to explain why they did so. Out of the 330 answers, 52 positive replies and 45 negative ones were counted.

First, considering the 52 "Yes" answers - or 15.7% - 90% of respondents indicated which vocabularies they used. Ten entries were available for them to complete. T7.4 observed that the number of answers declines entry after entry.

T7.4 categorised the results per entry as shown in Annex 3 in order to analyse if some answers could be highlighted - for instance, if one vocabulary was largely identified right from the first entry by a majority of researchers. A list of 159 answers was gathered. The analysis revealed that none of the listed vocabularies was widely used by the SSH research community. For instance, for the first entry, 47 answers were submitted in total, out of which "Pactols" is mentioned by 4 respondents, followed by "DDI Vocabularies", "Dublin Core" and "ELSST" with 2 repetitions each.

The vocabularies that were most often mentioned by the respondents were:

- DDI with a total of 9 repetitions
- Getty with a total of 5 repetitions
- CESSDA Controlled Vocabularies with a total of 4 repetitions
- ELSST with a total of 4 repetitions
- Dublin Core with a total of 4 repetitions
- Pactols with a total of 4 repetitions.

Out of the 72 participants who provided complete answers, 45 indicated that they did not use vocabularies. Based on the 38 explanations provided, the following reasons have been identified:

- respondents did not know what a vocabulary was used for;
- respondents did not know what its purpose in the light of their research project was;
- respondents thought it was too time-consuming to use vocabularies or to learn their purpose.

On the areas, discipline specificity and languages of vocabularies

For this fifth question, respondents could select multiple answers.

All answers considered and regarding the respondents who answered positively to the fourth question, here are the results obtained:

- 45.3% for Named entities: persons
- 52. 8% for Named entities: geographical entities
- 28.3% for Named entities: institutions
- 45.3% for Concepts: Scholarly activities
- 58.5% for Concepts: Disciplines
- 54.7% for Concepts: General concepts
- 30.2% for Others: please specify.

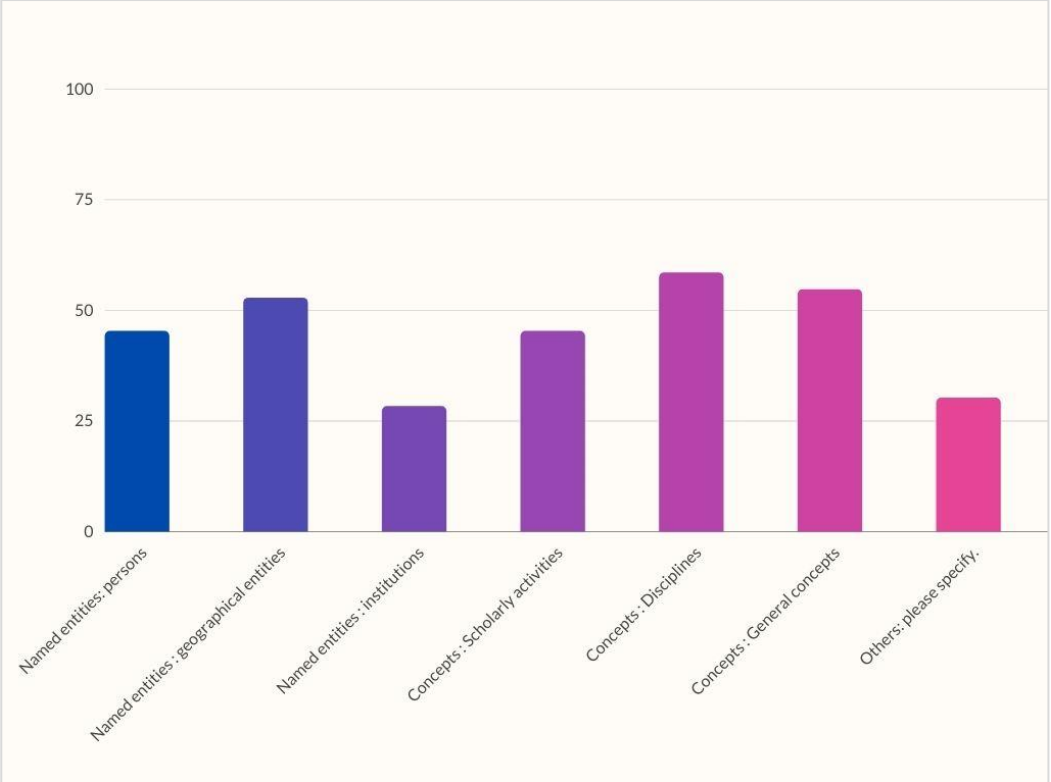

Figure 3. Vocabularies' categories' types all answered considered

For the "Others" category, here are the listed results: named entities; language names; research data; specific concepts; words describing social interactions; cultural heritage; chronological and functional definitions; subjects; morphemic glosses; archaeological types of objects; metadata; to determine the scope of a word; bibliographical data, and EU terminology.

For the sixth question ("In which languages is/are this/those vocabularies available in?"), about 90.57% of the 52 respondents provided an answer. Top three answers are (1) English with 30 repetitions, (2) French with 24 repetitions and (3) German and Spanish joint with 6 repetitions each.

Question number seven ("Are/is the vocabulary(ies) discipline-specific?") gathered the following answers:

- Yes for 77.4%
- No for 15%
- No answer for 7.6%.

On vocabularies' alignments

All complete answers considered and regarding the respondents that answered positively to the fourth question, here are the results obtained to question number eight ("Is/are the vocabulary(ies) you are using

aligned with other ones?"). Out of the total of respondents who use vocabularies, 64.3% indicate that their vocabularies are aligned with other ones, 14.3% with none, and 21.4% don't know.

Regarding the "yes" respondents, all specified the vocabularies with which they are aligned – see Annex 4. Here are the top three answers for alignment: (1) Getty, (2) Wikidata and (3) Dublin Core. The respondents who did not align their vocabularies expressed a "*lack of quality and time*", or an absence of the need to do so.

Question 9 ("With which vocabulary(ies) would you like the one(s) you're using to be aligned with? And why?") pursued the questioning on the alignment of vocabularies in SSH. For this question, out of the 52 vocabularies' users, 88.7% (all answers considered) provided an answer – see Annex 5. Results indicate that respondents wish for the vocabulary(ies) they are using to be aligned with Getty and Wikidata. Respondents also mentioned their wish to see alignment in terms of languages.

On the subject of languages, the following question ("In which language(s) would you like the vocabulary that you are using to be employed in? And why?") further scrutinises the alignment practices. Respondents wished their vocabularies to be available and aligned in the following languages:

- English with 20 repetitions,
- French with 14 repetitions,
- Spanish with 7 repetitions.

On vocabularies' obsolescence, self-creation and online availability

When asked whether the vocabularies used may be outdated, 13.2% answered *yes* to the eleventh question[13], while 75.5% answered *no* and 11.3% did not answer. Question 12 inquired about the use of self-created vocabularies and about the frequency of their updates. The 45.3% of "yes" respondents provided various answers regarding the update:

- Some considered updating the vocabularies on an irregular basis, only "*when necessary.*"
- Others never did it, either because of a lack of will or because the vocabulary had just been created.
- Some respondents updated their vocabularies regularly: "*twice a year*", "*each month*", "*regularly*", "*always*".

Finally, the last question of the survey ("Are the vocabularies available online? Are they available in a machine-readable format?") gathered the following results. To the first part of the question:

- 69.8% answered yes and provided the URL when possible – see Annex 6
- 20.8% answered no
- 9.4% didn't provide an answer.

[13] Considering all answers and "YES" answers' respondents to question 4.

On the machine-readable structured format of those:

- 64.2% answered yes
- 5.7% no
- 30.1% didn't provide an answer.

Concluding remarks

This section summarises and discusses the results of the survey.

Regarding the respondents' profiles, T7.4 notes that two disciplines systematically came out: **Linguistics** and **Archaeology & Prehistory**. Two others also stood out: History and Sociology. Finally, it is also worth mentioning that considering *yes* answers to the fourth question "Do you/your organisation use any vocabulary(ies)?" two new disciplines appear: History, Philosophy and Sociology of Science and Art & Art History. Future work on vocabularies could target researchers from these four disciplines to learn more about their needs in terms of vocabularies.

As for the organisation type, three categories stood out: **Universities and Research Performing Organisations**, **Researchers** and **Research Libraries & Archives**. Also, the most mentioned organisation name was: **CNRS**. Finally, regarding respondents' country of residence, most reside in **France**. Because of the large representation of French residents, it induces a bias regarding the survey's results. Future studies should consider this and attempt to represent more equally other nationalities.

The second part of the survey reveals that the respondents use the six following vocabularies: **DDI**, **Getty**, **CESSDA Controlled Vocabularies**, **ELSST**, **Dublin Core** and **Pactols**. However, due to the insufficient response rate, it is not possible to conclude that one vocabulary is particularly used within the SSH research community.

The diversity of the answers teaches that the notion of "vocabulary" is very broad and that vocabularies could be categorised as follows:

1) Metadata schemas (Dublin Core)
2) Generic SSH thesaurus (ELSST)
3) Specialised thesaurus in SSH fields (Pactols)
4) Dictionaries and glossaries
5) Ontologies representing and structuring data (SKOS).
6) Gazetteers, i.e. authority files of particulars, persons, places (Geonames, Getty TGN)
7) Research identifiers (ORCID).

This lack of common understanding - based on the rather broad definition the term conveys - is also reflected in respondents' *no* answers' justification: "*I don't know what a vocabulary is*" came out a few times.

In the light of the vocabularies' purposes and the area covered, they mainly cover three topics for respondents: **Geographical entities**, **Disciplines** and **General concepts**.

Also, 77.4% of respondents consider those **vocabularies as discipline-specific**. This is illustrated by the named vocabularies with regard to the listed disciplines. Getty and Pactols are specialized thesauri related to Archaeology and History corresponding to well-represented disciplines among the answers. Also, a connection can be made between the metadata description language (DC and DDI) or SKOS and the type of organization like Research Libraries and EOSC thematic, which supports Information System, Open Science and working on Semantic Web and the computational linguistic fields.

In the light of the language of the vocabularies commonly used by respondents, the results indicate that they are mostly available in **English**, **French** and **German** and **Spanish** (both equally ranked). In addition, most respondents wish for an alignment of their vocabularies with other languages: **English**, **French** and **Spanish**. However, one needs to be reminded of the French bias of this survey: with a majority of French residents' respondents, the "French" language result can be expected.

Considering vocabularies' alignment, the survey highlights that **64.3% of respondents practice vocabularies aligned** with other ones, especially **Getty**, **Wikidata** and **Dublin Core**. Also, **88.7% wish for an alignment with other vocabularies**, especially **Getty** and **Wikidata**. The answers on alignments like Getty, Wikidata and DC, show that there is a demand more related to knowledge structuring, with a machine-readable aspect, open and interoperable. Getty also respects the Linked Open Data criteria[14].

Finally, in the light of vocabularies potential outdatedness and online availability, **13.2% of respondents considered their vocabularies to be outdated against 75.5% of up-to-dateness**. In addition, 45.3% of respondents explained they were using self-created vocabularies. Also, **69.8% indicated that the vocabularies were available online** and **64.2% that they were available in a machine-readable format**.

[14]The Getty Research Institute. Getty Vocabularies as Linked Open Data https://www.getty.edu/research/tools/vocabularies/lod/index.html, accessed on December 18th, 2020

4. Relations to the SSHOC project

WP7's work on the SSH Open Marketplace mapping process

As explained in the introduction, vocabularies are crucial for the SSH Open Marketplace as a discovery platform. However, given that most of the content is being ingested from existing external sources, T7.4 is faced with the additional challenge of "local" vocabularies used at these sources and the need to harmonize or map them to the concepts of vocabularies adopted within the Marketplace.

This is a rather complex task, especially because it involves intertwining machine and human action. During the ingestion process, values from third parties sources can be automatically identified and ingested in the system, but the mapping itself - i.e. matching values from the ingested sources to the existing concepts structured as vocabularies in the SSH Open Marketplace - sometimes requires a human decision. If the system is not able to identify a new value as synonym to an existing concept, an editor/moderator needs to assess the meaning of the new term and either to manually add it as a synonym to an existing concept or to create a new concept in the given vocabulary. The system specification D7.1 elaborates on the mechanism of vocabulary management. The main idea is to create so-called candidate concepts during the ingestion process, which can be used provisionally immediately but require a review and approval by a human editor in a subsequent step.

In any case conceptually, and formally the SKOS data model lends itself well to capture the information needed for the mapping process: it allows recording multiple alternative labels (skos:altLabel) for a concept, which can be considered as lexical synonyms, and it also allows expressing semantic relationships between concepts (broader, narrower) even between concepts from different vocabularies.

Regarding the intended purpose of the survey as a source for vocabularies for describing items in the Marketplace, most of the vocabularies gathered in the survey are too specific to be applicable. Additionally, the respondents often named schemas (i.e. sets of fields) as vocabularies, information that can potentially be useful when defining mappings between the Marketplace data model and these schemas either for import or export purposes, but it is not usable for defining specific controlled vocabularies for individual fields. However, a small set of vocabularies from the survey indeed is relevant for the Marketplace, for example[15]:

- TaDiRAH - Taxonomy of Research Activities in the Humanities
- ISO-639 - Language codes
- ELSST - European Language Social Science Thesaurus
- DYAS Humanities Thesaurus

[15] In general, vocabularies that could potentially be used in the Marketplace as described in D7.1 were almost all mentioned in the answer 4 of the survey. Only CLAVAS, the PICO Thesaurus, the IANA mime/type and NeDiMAH were not mentioned at all.

Yet another kind of "vocabularies" appears in the survey responses, which T7.4 wants to refer to as "gazetteers", or authority files of particulars, i.e. persons or places, like ORCID, Geonames or TGN - Getty Thesaurus of Geographic Names. These will potentially become relevant for the Marketplace as an external identifier when referring to specific entities. Foremost, ORCID is already foreseen as the primary external identifier for persons (both contributors to the Marketplace and authors or creators of the items described in the Marketplace).

T7.4 wants to highlight two issues arising from the reuse of vocabularies having to do with inherent problems of semantic interpretation of lexical terms for describing concepts and conceptualisation of a given domain of discourse:

a) On the one hand, the **scope of a vocabulary** may be too narrow or too broad for the specific purpose, i.e. important concepts may be missing, or there may be too many concepts that are not relevant. This can be tackled by customising an existing vocabulary, if possible, by proposing changes (esp. addition of new concepts) directly to the authoritative source of the vocabulary. Alternatively, a derivative vocabulary can be devised under the control of the repurposing agent (in our case the Marketplace editorial team), where any needed changes can be performed freely, however at the cost of potentially limiting the semantic interoperability and adding to the proliferation of vocabularies (ending up with a separate vocabulary for each context or purpose).
b) On the other hand, there may be "false friends, i.e. concepts or terms that sound fitting lexically, but actually **mean something different** in the original context, compared to the interpretation in the new context. While this is a general problem inherent to communication, semiotics and interpretation of symbols, one approach to remedy is to make the semantic explicit by means of definitions, explanations and examples that should be an integral part of a well-curated vocabulary.

In case the identified vocabularies are not suitable to the description of items in the Marketplace, but widely used by some communities and well-curated, an **entry will be created in the Marketplace**. The idea is to give visibility to this resource and also to be able to liaise with other entries in the Marketplace.

As for other Marketplace entries, it will be necessary to check regularly the evolution and the degree of use of these vocabularies. The life cycle of this type of resource is quite different from that of a tool and works over a long period of times.

Regarding the selection criteria, in addition to those mentioned above which are oriented towards scientific relevance, the use of standardized technologies (e.g. RDF) to publish them would be, of course, important criteria as well as the user licence.

SSHOC's WP3 activities on vocabularies' use consolidation

One of the main goals of the SSHOC Task 3.1 Multilingual Terminology, as described in the work plan of WP3, is to find a suitable vocabulary server and publication platform to increase the accessibility and improve discovery by non-native speakers. In the early stages of the SSHOC project, WP3.1 started with the collection

of the SSHOC requirements for the vocabulary registry which were recorded in a milestone report (SSHOC Milestone 8 – SSHOC Considerations for vocabulary platforms.) After identifying some overlap of vocabulary activities across work packages, CLARIN ERIC took the initiative of harmonizing the efforts as part of their SSHOC Task 3.1

Part of the work plan is to collect, register and harmonize the most "known" SSH controlled vocabularies, taxonomies and thesauri, and find a sustainable vocabulary platform to host and publish them. To achieve this goal, a series of online information sessions[16] and a virtual workshop[17] have been recently organized, where experts and users of vocabulary platforms discussed the requirements and exchanged best practices. The insights will be used to reassess the preliminary recommendations of the milestone report and to further investigate how the harmonisation of the SSHOC vocabularies can be added to the SSHOC outputs, such as the Marketplace.

Links with other SSHOC initiatives: the SSHOC Reference Ontology (SSHOCro)

Task 7 from Work Package 4 aims to define the SSHOC Reference Ontology (SSHOCro)[18]. SSHOCro proposes an ontological model and RDF schema to be used as a top-level ontology for organizing knowledge and information found distributed across various primary sources of information in the Social Sciences and Humanities Open Cloud.

SSHOCro is a common meta-level schema to provide a semantic interoperability framework for the description of the data life cycle used by Social Science and Humanities researchers. SSHOCro is modelled as an extension of CIDOC CRM, the ISO standard ontology for Cultural Heritage data, from which it inherits its event-centric orientation.

In practical terms, the use of such a model and schema for the research community is twofold: it can be applied as a standard to be used in the step of devising and implementing a metadata capture scheme for tracking the data lifecycle in individual projects, institutions and disciplines; it can also be used to map, transform and integrate existing data across projects, institutions and disciplines into interoperable pools of information for reuse and exploitation. In this context, keeping track of the processes involved in the data lifecycle amounts to associating each stage to a set of activities performed in it.

[16] SSHOC Online Information Sessions: Open-Source Vocabulary Hosting and Management Platforms: https://www.clarin.eu/event/2020/sshoc-online-information-sessions-open-source-vocabulary-hosting-and-management; accessed 18 December 2020.

[17] Virtual workshop: https://www.clarin.eu/event/2020/sshoc-considerations-vocabulary-platforms-virtual-workshop, accessed 18 December 2020.

[18] See Bekiari,Chrysoula, Kritsotaki, Athina, & Tsouloucha, Eleni. (2020). SSHOC D4.18 SSHOC Reference Ontology (beta version) (Version v1.0). Zenodo, https://doi.org/10.5281/zenodo.3744860 accessed 18 December 2020.

This framework will be essential for interoperability at a structural level which will be complementary to the interoperability provided by the harmonisation of the SSHOC vocabularies to map, transform and integrate existing data across projects, institutions and disciplines into interoperable pools of information for reuse and exploitation.

5. Conclusion

This document presents the results of a survey on vocabularies in SSH research practice that the SSHOC T7.4 team conducted in the first half of 2020. The main goal of this survey was to gain a better understanding of the actual knowledge and use of vocabularies in SSH research communities in order to inform and improve the SSH Open Marketplace developments (WP7). The secondary objectives were to complement related work conducted in SSHOC WP3 and the TRIPLE project. In order to achieve these goals, the T7.4 team developed a survey with 14 questions regarding vocabulary usage and work practices. The survey was available for 4 months among the SSH research communities. Out of 330 answers received, 72 were complete and 42 proved to be fully relevant.

Even if the survey results are inconclusive, several points can be highlighted in the light of this deliverable's threefold objective. First, this study can be seen as a contribution to the analysis of SSH research practices' when it comes to the use of vocabularies. Although it is not possible to clearly identify specific trends and practices within the community, the answers of the 42 participants who indicated that their organisations used vocabularies demonstrate that the SSH community's apprehension of vocabularies is very disparate and heterogeneous.

This point has to be understood through the very definition of what a vocabulary represents for the community itself. The concept remains rather broad and perhaps unclear to them, as it is reflected by the "ORCID" reply - a research identifier. This very heterogeneity made it difficult to identify which vocabularies are suitable for the SSH Open Marketplace's content description: only a small subset of the vocabularies mentioned in the survey answers (TaDIRAH, ISO-639, ELSST, DYAS) can be reused.

Nevertheless, the insights collected through this survey are relevant for future studies: the most used and well-curated SSH vocabularies identified via this survey could be added as resources in the Marketplace, while CLARIN and SSHOC could further identify and investigate the "orphan" vocabularies mentioned by the respondents. These ''orphan'' vocabularies need to be registered and preserved in the SSH community. Future extensive work towards the SSH community on vocabularies is currently undertaken by CLARIN ERIC and WP3. Through this study, T7.4 contributes to informing future work on the consolidation of the use of vocabularies in the SSH communities.

6. References

Barbot, Laure & Moranville Yoann & Fischer, Frank & Petitfils, Clara & Ďurčo, Matej et al. *D7.1 System Specification - SSH Open Marketplace*, 2019. https://zenodo.org/record/3547649#.X8oQaqpKhQJ

Broeder, Daan & Trippel, Thorsten & Degl'Innocenti, Emiliano & Giacomi, Roberta & Sanesi, Maurizio et al. *D3.1 Report on SSHOC (meta)data interoperability problems*, 2019. https://zenodo.org/record/3569868#.X8oRyKpKhQJ

European Commission. "Mapping of Research in European social sciences and humanities". *CORDIS website*, 2005. https://cordis.europa.eu/project/id/HPSE-CT-2002-60060

Monachini, Monica & Jääskeläinen, Taina. *M3.1 Choice of Vocabulary publication platform for SSHOC*, 2019.

GOFAIR. "FAIRification Process" GOFAIR website, unknown. https://www.go-fair.org/fair-principles/fairification-process/

Hider, Philip. "Vocabularies". *Information Resources Description: Creating and managing metadata*, 2012. https://books.google.de/books?id=jGpVPgAACAAJ&lpg=PP1&pg=PA151#v=onepage&q&f=false

Schwabe, Annika & Ausserhofer, Julian & Marino, Leonardo & Willems, Marieke & Muscella, Silvana et al. *D2.1 SSHOC Overall Communication and Outreach Plan*, 2019. https://www.sshopencloud.eu/sites/default/files/D2.1_SSHOC_Overall_Communication_Outreach_Plan_v1.0.pdf

Torma, Martina & Kalaitzi, Vasso & Vipavc Brvar, Irena & Fišer, Darja & Pahor de Maiti et al. *D6.1 Community Engagement Strategy*, 2019. https://zenodo.org/record/3592244#.X8oTIapKhQI

Annex 1 - The questionnaire

Profile

1. What is your social sciences and humanities discipline?

- Archaeology and Prehistory
- Architecture, Space Management
- Art and Art History
- Biological Anthropology
- Classical Studies
- Cultural Heritage and Museology
- Demography
- Economics and Finances
- Education
- Environmental Studies
- Gender Studies
- Geography
- History
- History, Philosophy and Sociology of Sciences
- Law
- Linguistics
- Literature
- Management
- Methods and Statistics
- Musicology and performing arts
- Philosophy
- Political Science
- Psychology
- Religions
- Social Anthropology
- Sociology

2.1. What is the name of your research organisation infrastructure?

2.2. How would you describe you/your organisation?

- Research Libraries and Archives
- Research & e-Infrastructures & EOSC thematic clusters
- Universities and Research Performing Organisations
- Research Funding Organisations

- Policy-Making Organisations
- Private Sectors and Industry Players
- Civil Society and Citizen Scientists
- Others.

3. What is your country of residence? (Selecting among the list of countries)

On the use of vocabularies

4. Do you/your organisation use vocabularies (both digital and/or printed ones)?

- Yes :
    - If yes, which ones ? (10 entries available)
- No :
    - If no, why?

5. What is/are the vocabulary(ies) used for ? Which area/topic is the vocabulary covering?

- Named entities: persons
- Named entities: geographical entities
- Named entities: institutions
- Concepts : Scholarly activities
- Concepts : Disciplines
- Concepts : General concepts
- Others : please specify

6. In which languages is/are this/that vocabulary(ies) available in?

7. Are/is the vocabulary(ies) discipline-specific ?

- Yes / Yes, some
- No

8. Is/are the vocabulary(ies) you're using aligned with other ones?

- Yes / Yes, some
    - If yes, which ones?
- No :
    - If no, why?
- I don't know

9. With which vocabulary(ies) would you like the one(s) you're using to be aligned with ? And why?

10. In which language(s) would you like the vocabulary that you're using to be employed in?

10.1 And why?

11. Are the vocabularies that you are using outdated?

- Yes
- No

11.1 If they are, why?

12. Are you using self-created vocabularies?

- Yes
- No

12.1 If yes, how often do you update them?

13. Are the vocabularies available online?

- Yes / Yes, some
- No

13.1 Please, specify the URL

13.2 Are they available in a machine-readable structured format?

- Yes / Yes, some
- No

End of survey

14. Would you like to take part in a follow-up group? If you agree you will have the chance to test and provide advice on vocabularies

Subscribe to the SSHOC Newsletter Membership of the community entitles you to receive regular news updates and progress reports

- I wish to subscribe
- Please insert a valid email address.

Annex 2 - Respondents' research organisation / infrastructure

Question 2.1 "What is the name of your research organisation/infrastructure?". List of respondents' affiliation with research organisation and number of repetitions :

BELGIUM :

- KU Leuven University : 1

BRAZIL :

- Federal University of the State of Rio de Janeiro : 1

FINLAND :

- Tampere University: 1

FRANCE :

- Centre National de la Recherche Scientifique : 18
- Ecole des Ponts: 1
- Paris University : 1 + 6
    - Cergy-Pontoise University : 1
    - Paris Nanterre University : 3
    - Paris 8 University: 1
    - Sorbonne Nouvelle University : 1
- Sciences Po: 1
- University of Caen Normandie : 1
- University of Lille : 1

GERMANY :

- Bavarian State Library : 1

GREECE :

- Academy of Athens : 2

THE NETHERLANDS :

- Utrecht University : 1

NORWAY :

- UiT The Arctic University of Norway : 1

POLAND :

- University of Wroclaw : 1

SLOVENIA :

- Jozef Stefan Institute : 1

SPAIN :

- Spanish National Research Council : 1
- University of Grenada : 1
- University of Valencia : 1

SWEDEN :

- Sweden National Data Service : 1

SWITZERLAND :

- Swiss Institute of Bioinformatics : 1

UNITED KINGDOM :

- UK Data Service : 1
- University of Essex : 1
- University of London - School of Advanced Study : 1
- University of South Wales : 1

UNITED STATES OF AMERICA :

- Kenyon College : 1

OTHERS :

- KNAW/HuC : 1

Annex 3 - On the use of SSH vocabularies

Question 4 "Do you/your organisation use vocabularies (both digital and printed and/or printed ones)?"

VOCABULARIES	ENTRY #1	ENTRY #2	ENTRY #3	ENTRY #4	ENTRY #5	ENTRY #6	ENTRY #7	ENTRY #8	ENTRY #9	ENTRY #10	TOTAL
DDI Alliance Controlled Vocabularies : AnalysisUnit / DDI 2 / DDI Alliance CV: ModeOfCollection / DDI 3 / DDI Alliance Controlled Vocabularies: ResponseUnit / DDI Alliance Controlled Vocabularies: TimeMethod / DDI Controlled Vocabularies / DDI Alliance Controlled Vocabularies: SamplingProcedure	2	2	2	1	1	1		1			9
ISO 639 Language Codes / ISO 3166 country codes / PN-ISO 5127:2005 Informacja i dokumentacja – Terminologia	1				1			1			3
Leipzig Glossing Rules		1									1
ELSST	2	1				1					4
HASSET	1	1									2
Rameau	1	1									2
FAIRisation	1										1
TEI (Text encoding initiative)	1	1									2
dcterms	1										1
Dublin Core categories / / dce-Dublin Core Metadata Element Set	2	1	1								4
Thésaurus Savoirs	1										1
GeoEthno	1										1
Common French dictionnary	1										1
TheA - Thesaurus-Antiquité	1										1
Bibliographic Ontology	1										1
Pactols	4										4
Thesaurus Patriarche	1										1
IdRef	1	1									2
Vocabulaire de psychosociologie	1										1
Occupation_API with 4,000 occupational titles for 100+ locales (coded ISCO08) (locale = language by country)	1										1
OLAC metadata schemes	1										1
didactique des langues	1										1
archaeological periods	1										1

DYAS HUMANITIES THESAURUS	1										1
Thesaulangue	1										1
coar	1										1
Cybergeo keyword list	1										1
Gemeinsame Normdatei (GND)	1										1
heritage data	1										1
Encyklopedia wiedzy o książce, pod red. A. Birkenmajera, B. Kocowskiego, J. Trzynadlowskiego, Wrocław 1971.	1										1
Thesaurus of the Bibliography of British and Irish History	1										1
BIBO the bibliographic ontology	1			1	1						3
GTAA	1										1
vocabularies related to CMDI metadataset	1										1
UNESCO Thesaurus	1										1
Getty AAT / Getty CONA / Getty TGN / Getty Art & Architecture Thesaurus® Online	1	1	1	1						1	5
Lexicon of Linguityics	1										1
TaDiRAH - Taxonomy of Digital Research Activities in the Humanities	1										1
opentheso	1										1
metadonnées		1									1
EAD		1									1
foaf (Friend of a Friend)		1	1		1						3
Dictionnaire de l'Ethnologie et de l'Anthropologie (Bonte & Izard ed., Puf)		1									1
extracted terms with test-mining tools	1										1
IATE Terminology Database	1										1
Vocabulaire de la psychanalyse		1									1
Region_API with over 20,000 provinces and cities for 100+ locales (for EU coded NUTS)		1									1
Français sur objectif universitaire		1									1
Usage		1									1

crm- CIDOC Conceptual Reference Model		1									1
Terminologie des sciences du langue INIST		1									1
openaire		1									1
HAL nomenclature of SSH productions		1									1
MeSH		1									1
BSB-DDC		1									1
Encyklopedia książki, pod red. A. Żbikowskiej-Migoń, M. Skalskiej-Zlat, t. 1–2, Wrocław 2017.		1									1
Library of Congress Subject Headings		1									1
ONTOLEX-lemon Lexical Data RDF model / The RDF Concepts Vocabulary		1			1						2
Tesauro de Ciencias de la Documentación		1									1
The textiles		1									1
GOLD Ontology		1									1
World-Historical Gazetteer		1									1
ORCID			1								1
ecdotique			1								1
bio			1								2
Musical Instruments (internal voc)			1								1
Dictionnaire de sociologie clinique			1								1
Cost-of-Living_API with over 100 food items for 100+ locales			1								1
français langue étrangère			1								1
dbpedia-owl- The DBpedia Ontology			1								1
Termsciences INIST			1								1
JEL			1								1
fundref			1								1
SH3 list from ERC			1								1
GEMET			1								1
Thesaurus Sozialwissenschaft			1								1

Encyklopedia współczesnego bibliotekarstwa polskiego, pod red. K. Głombiowskiego, B. Świderskiego, H. Więckowskiej, Wrocław 1976.			1							1
Virtual International Authority File			1							1
LexDo Domain Ontology of Lexicography			1							1
European Education Thesaurus			1							1
European Education Thesaurus			1							1
CESSDA Controlled Vocabularies / CESSDA Topics Classification				1	1				2	4
Audiovisuals carriers type (internal voc)					1					1
GeoNames				1	1	1				3
Publicionnaire				1						1
Industry_API with 400 industry categories for 100+ locales (coded NACrev2.0)					1					1
Refnum					1					1
lexique scientifique					1					1
iDAI Gazetteer https://gazetteer.dainst.org/app/#!/home					1					1
agrovoc					1					1
Sections of the French National Committee of Universities					1					1
EnvThes					1					1
iDAI-thesauri					1					1
Słownik terminologiczny informacji naukowej, pod red. M. Dembowskiej, Wrocław 1979.					1					1
ODLIS Online Dictionary of Library and Information Science					1					1
CLARIN Concept Registry			1							1
skos				1			1			2
voice description (internal voc)				1						1
lexinfo				1						1
Les 100 mots de la sociologie				1						1

Religion_API with up to 10 religious categories for 100+locales (ready by end 2020)				1							1
lexique transdisciplinaire				1							1
Grand dictionnaire de terminologie				1							1
dewey				1							1
loc				1							1
sections of the National Committee of the CNRS				1							1
GCMD				1							1
Standard-Thesaurus Wirtschaft				1							1
Podręczny słownik bibliotekarza, oprac. G. Czapnik, Z. Gruszka przy współudz. H. Tadeusiewicz, Warszawa 2011.				1							1
PROV-O				1							1
International Labour Organization ILO Thesaurus				1							1
Musical genres (internal voc)						1					1
IconClass						1					1
Web Annotation Ontology						1					1
Les termes clés de l'analyse du discours						1					1
Education_API, for 100 countries, see https://www.surveycodings.org/levels-education						1					1
Leipzig Glossing Rules (for interlinear glosses)						1					1
émotions						1					1
Dictionnaire méthodique de l'architecture grecque et romaine I, II et III (Ginouvès et Martin, 1985, 1992, 1998)						1					2
Klassifikation Sozialwissenschaften						1					1
Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych, oprac. B. Bojar, Warszawa 2002.						1					1
LEXVO						1					1

Glossary of Library & Information Science						1					1
viaf							2				2
Key words (internal voc)							1				1
Dictionnaire de la fatigue							1				2
Collective Agreements Database, with over 1000 Agreements from some 40 countries, full text and coded, with keywords							1				1
notices d'autorité BnF							1				1
sociolinguistique							1				1
Ornement - Vocabulaire typologique et technique, Laurence de Finance, Pascal Liévaux Collection Vocabulaires, éditions du Patrimoine							1				1
INSPIRE data themes / INSPIRE Topic categories							1			1	2
JEL Classification							1				1
Sternik. Słownik terminologiczny z zakresu bibliografii i katalogowania, pod red. J. Packa, http: //sternik.bn.org. pl/vocab/index.php.							1				1
LCSH Library of Congress Subject Headings							1				1
Tesauro de Biblioteconomía y Documentación							1				1
Phonographic publishers (internal voc)								1			1
Ontology for Media Resources								1			1
Dictionnaire des risques psychosociaux								1			1
Global Labour Law Database, with labour law for 100+ countries, full text (rephrased for popular use) and coded								1			1
psycholinguistique								1			1

Sculpture - Méthode et vocabulaire, Marie-Thérèse Baudry - avec la collaboration de Dominique Bozo Collection Vocabulaires, éditions du Patrimoine								1			1
The Swedish standard classification of fields of research								1			1
Tesauro ERIC								1			1
Phonographic collections (internal voc)									1		1
Information Economic Meta Language									1		1
Minimum Wages Database, with specified MW rates, for more than 120 countries									1		1
lexicologie sémantique									1		1
Espace urbain - Vocabulaire et morphologie, Bernard Gauthiez Collection Vocabulaires, editions du Patrimoine									1		1
Encyklopedia wiedzy o prasie, pod red. J. Maślanki, Wrocław 1976									1		1
UNBIS Thesaurus									1		1
Finnish classification of scientific disciplines										1	1
Populations (internal voc)										1	1
Stato Fr (OBO)										1	1
Living Wages Database, for 75 countries										1	1
littéracies universitaires										1	1
Joan M. Reitz, ODLIS. Online dictionary for library and information science, https://products.abc-clio.com/ODLIS/odlis_about.aspx.										1	1
Tesauro Brasileiro de Ciência da Informação										1	1
No answer	5	19	28	32	34	37	37	41	43	43	319

Annex 4 - Vocabularies' alignment

Question 8 "Is/are the vocabulary(ies) you're using aligned with other ones? Which ones?"

Vocabularies aligned with	Total/Repetitions
Dublin Core	3
DDI Alliance	1
Eurovoc	1
LCSH	1
BNE	1
Linguistics	1
XML	1
VIAF	1
ISNI	1
Nakala categories	1
TheA	1
SKOS	1
Web Annotation Ontology	1
Information Economic Meta Language	1
OpenTheso	1
Vocabularies are coded according to standard classifications, such as ISCO08 for occupations, NACE for industries, ISCED for educations, and NUTS for geographical regions	1
inventories of interlinear glosses aligned with Leipzig Glossing Rules, in particular	1
english for academic purposes, university litteracies...	1
Wikidata	4
Getty (AAT)	5
IDai Gazetteer	1
Pleiades	1
Insee	1

Art & Architecture thesaurus	1
Geonames	1
IdRef	1
Rameau	1
Pactols	1
openaire	1
SH3 of ERC	1
OpenEdition platform	1
Informal group of journal of geography	1
GCMD	1
GEMET	1
EnvThes	1
GND	1
BBS-DDC	1
Thesaurus Sozialwissenschaft	1
Linked Open Data Cloud	1
daf	1
UNESCO Thesaurus	1
ILO Thesaurus	1
UNBIS Thesaurus	1

Annex 5 - Expectations for vocabularies' alignment

Question 9 "With which vocabularies would you like the one(s) you are using to be aligned with?"

Answers	Total
ORCID to be aligned with ROR	1
EuroVoc thesaurus	1
ICPSR thesaurus	2
Linguistics	1
Wikidata	3
Autoalignement with Viaf	1
isni	2
have categories the most broad and internationally extended, in order to allow harvesting of data and interoperability of data	1
GeoNames	1
IconClass	1
Getty (AAT)	4
Rameau	1
Information Economic Meta Language	1
Trismegistos	1
Pleiades	1
Vocabularies applied to historical research	1
It would be desirable to have better alignment between vocabularies for glossing interlinearized text. This raises epistemological difficulties but a first-pass equivalence (and more elaborate forms of alignment, specifying degree of closeness, for instance) would be of great use	1
BNF	1
Opentheso	1
Pactols	1
Periodo	1
JEL	1
loc to be aligned with europeana	1
vocabularies in English/other languages	3
ADFD	1
CVVBCVB	1
Unesco Thesaurus	1
GOLD ontology	1
QSG	1
No opinion / no / not sure to understand	13

Annex 6 - Online vocabularies' URLs

Question 13 "Are the vocabularies available online?"

List of URLs mentioned by respondents:

- American Economic Association - JEL Classification system: https://www.aeaweb.org/econlit/jelCodes.php
- BIO - A vocabulary for biographical information: https://vocab.org/bio/
- BNF - Référentiel d'enrichissement des métadonnées version Refnum: https://www.bnf.fr/sites/default/files/2018-11/ref_num_metadonnees_refnum3.pdf
- BNF Catalogue Général: https://catalogue.bnf.fr/recherche-autorite.do
- BNF Rameau: https://rameau.bnf.fr
- BVB - Bibliotheks Verbund Bayern: https://www.bib-bvb.de/web/b3kat/klassifikatorische-sacherschlie-ung-im-b3kat
- CESSDA Data Catalogue: https://datacatalogue.cessda.eu/
- CESSDA Vocabulary Service:https://vocabularies.cessda.eu/#!detail/TopicClassification?url=https%3A%2F%2Fvocabularies.cessda.eu%2FTopicClassification%2Fen%2F3.0
- CLARIN Concept Registry Browser: https://concepts.clarin.eu/ccr/browser/
- COAR - Confederation of Open Access Repositories: https://www.coar-repositories.org/news-updates/what-we-do/controlled-vocabularies/
- CREM - Centre de Recherche en Ethnomusicologie: https://archives.crem-cnrs.fr/instruments/
- Crossref: https://www.crossref.org/dashboard/
- Datu Thésaurus Savoirs: https://datu.ehess.fr/savoirs/
- Dictionnaire de la fatigue: https://www.cairn.info/dictionnaire-de-la-fatigue--9782600047135.htm
- Dictionnaire de sociologie clinique: https://www.cairn.info/dictionnaire-de-sociologie-clinique--9782749257648.htm
- Dublin Core Metadata Initiative: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
- DYAS Humanities Thesaurus: https://humanitiesthesaurus.academyofathens.gr/
- EAD - Encoding Archival Description: https://francearchives.fr/file/0def64f5a10f3f1ae03fdea59399a3e0755ef157/static_1066.pdf
- ELSST - European Language Social Science Thesaurus: https://elsst.ukdataservice.ac.uk/
- EMM Survey Registry: https://ethmigsurveydatahub.eu/emmregistry/
- ERIC: https://eric.ed.gov/?ti=all
- European Education Thesaurus: https://vocabularyserver.com/tee/en/
- FAO - Agrovoc: https://aims.fao.org/fr/agrovoc
- FOAF: http://xmlns.com/foaf/0.1/
- Geonames: https://www.geonames.org/

- Glossary of Library & Information Science: https://www.librarianshipstudies.com/2015/04/glossary-of-library-information-science.html
- GOLD: http://linguistics-ontology.org/
- HASSET : https://hasset.ukdataservice.ac.uk/
- Heritage Data : https://www.heritagedata.org/blog/vocabularies-provided/
- IdRef - Identifiants et référentiels pour l'enseignement supérieur et la recherche: https://www.idref.fr/
- ILO - International Labour Organisation: https://metadata.ilo.org/thesaurus.html
- ISO 5127:2017: www.iso.org/obp/ui/#!iso:std:59743:en
- ISO 639.2 "Codes for the Representation of Names of Languages": https://www.loc.gov/standards/iso639-2/php/code_list.php
- Les 100 mots de la sociologie: https://journals.openedition.org/sociologie/160
- LexDo: http://lexbib.org/lexdo
- Lexicon of Linguistics: https://lexicon.hum.uu.nl/
- Lexiscope: http://phraseotext.univ-grenoble-alpes.fr/lexicoscope/
- Library of Congress: http://id.loc.gov/authorities/subjects/sh85144057.html
- OCLC: https://www.oclc.org/en/dewey.html
- ODLIS - Online Dictionary for Library and Information Science: http://products.abc-clio.com/ODLIS/odlis_about.aspx
- OLAC Metadata: http://www.language-archives.org/OLAC/metadata.html
- OpenAIRE: https://www.openaire.eu/research-products-and-their-associated-types-in-openaire?highlight=WyJ2b2NhYnVsYXJ5Il0=
- OpenEdition Journals - cybergeo: https://journals.openedition.org/cybergeo/33
- Opentheso: https://thesaurus.mom.fr/opentheso/index.xhtml
- ORCID: https://orcid.org/
- ORTOLANG: https://www.ortolang.fr/market/terminologies/vocabulaire-de-sciences-du-langage/
- PACTOLs: https://pactols.frantiq.fr/opentheso/
- Pleiades: https://pleiades.stoa.org/
- Publictionnaire: http://publictionnaire.huma-num.fr/
- ScienQuest : https://corpora.aiakide.net/scientext20/
- Semantic scholar: https://www.semanticscholar.org/search?q=fran%C3%A7ais%20sur%20objectif%20universitaire&sort=relevance
- SKOS: https://ihr-webmaster.github.io/vocab-explorer/index.html
- SKOS: https://www.w3.org/2004/02/skos/
- SSHOC Survey Coding: https://www.surveycodings.org/
- TEI - Text Encoding Initiative : https://tei-c.org
- Tesauro Brasileiro de ciência da informação: http://sitehistorico.ibict.br/publicacoes-e-institucionais/tesauro-brasileiro-de-ciencia-da-informacao-1/copy_of_TESAUROCOMPLETOFINALCOMCAPA24102014.pdf

- Tesauro de biblioteconomia y documentacion: https://digital.csic.es/bitstream/10261/30255/1/Tesauro%20de%20Biblioteconomia%20y%20Documentacion.pdf
- Tesauro de ciencas de la documentacion: http://eprints.rclis.org/5875/1/TESAURO_DOCUTES.pdf
- The Bibliographic Ontology: http://bibliontology.com/
- The Getty Research Institute: https://www.getty.edu/research/tools/vocabularies/tgn/index.html
- The Leipzig Glossing Rules: https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf
- Trismegistos: https://www.trismegistos.org/index.php
- UK Data Service: https://beta.ukdataservice.ac.uk/
- UNESCO Thesaurus: http://vocabularies.unesco.org/browser/thesaurus/en/
- VIAF: http://viaf.org/
- Vocabulaire de psychologie: https://www.cairn.info/vocabulaire-de-psychosociologie--9782749229829.htm
- WageIndicator: https://wageindicator.org/